MATERIAL DIDÁCTICO
MATEMÁTICAS **7**

INICIACIÓN A LOS MÉTODOS NUMÉRICOS

José Antonio Ezquerro Fernández







INICIACIÓN A LOS MÉTODOS NUMÉRICOS

MATERIAL DIDÁCTICO

Matemáticas

nº 7

José Antonio Ezquerro Fernández

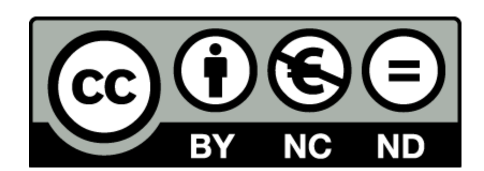
INICIACIÓN A LOS MÉTODOS NUMÉRICOS

UNIVERSIDAD DE LA RIOJA

SERVICIO DE PUBLICACIONES

2012

Ezquerro Fernández, José Antonio
Iniciación a los métodos numéricos / José Antonio Ezquerro Fernández. -
Logroño : Universidad de La Rioja, Servicio de Publicaciones, 2012.
144 p. ; 29 cm. (Material Didáctico. Matemáticas ; 7)
ISBN 978-84-695-2800-6
1. Análisis numérico. I. Universidad de La Rioja. Servicio de Publicaciones, ed.
519.6



publicaciones.unirioja.es

E-mail: publicaciones@unirioja.es

ISBN 978-84-695-2800-6

Edita: Universidad de La Rioja, Servicio de Publicaciones

A María

Prólogo

La búsqueda de soluciones reales ha cautivado la atención de los matemáticos desde sus primeros tiempos, ocupando un lugar importante en el estudio de las matemáticas. Encontrar una solución exacta de un problema puede llegar a ser imposible, o puede que no podamos encontrar una respuesta de forma conveniente en una gran cantidad de aplicaciones reales. Cuando esto ocurre, perseguiremos dar respuestas útiles que involucren la búsqueda de resultados aproximados suficientemente buenos. Esta es la razón de los *métodos numéricos*, teniendo muchos de ellos una larga historia.

Los métodos numéricos son técnicas matemáticas que se utilizan para resolver problemas matemáticos que no se pueden resolver, o que son difíciles de resolver, analíticamente. Una solución analítica es una solución exacta que tiene la forma de una expresión matemática en función de las variables asociadas al problema que se quiere resolver. Una solución numérica es un valor numérico aproximado (un número) de la solución. Aunque las soluciones numéricas son aproximaciones, pueden ser muy exactas. En muchos métodos numéricos los cálculos se ejecutan de manera iterativa hasta que se alcanza una exactitud deseada, de manera que tienen que ser lo suficientemente exactos como para satisfacer los requisitos de los problemas a resolver y lo suficientemente precisos para ser adecuados.

El origen de este texto es un curso introductorio de métodos numéricos que lleva impartiendo el autor en la Universidad de La Rioja desde el curso 2006-2007. El texto proporciona una introducción a los fundamentos más básicos de los métodos numéricos y a su utilidad para resolver problemas, ofreciendo una primera toma de contacto que sirva para conocer una parte del amplio catálogo que existe de métodos numéricos. El objetivo principal es poner a disposición de los estudiantes unas notas de fácil lectura con una colección de ejercicios que amplíen y refuercen lo que aprenden. Otras obras más completas, véanse las referencias bibliográficas, que contengan desarrollos teóricos y mayor número de ejemplos son un complemento ideal a este texto.

El texto va dirigido a estudiantes universitarios que van a utilizar los métodos numéricos por primera vez y que no tienen conocimientos previos de los mismos. Se presentan métodos numéricos elementales de una manera asequible y sin ser tratados de forma sistemática, de manera que este texto se pueda utilizar como guía inicial para un proceso posterior que profundice en los métodos numéricos de manera más detallada y extensa. A la hora de redactar el texto se ha tenido presente en todo momento a los estudiantes a los que está dirigido.

Los requisitos mínimos son el cálculo elemental, incluyendo los polinomios de Taylor, el álgebra matricial y tener nociones básicas de programación estructurada. Para los capítulos 6 y 7 es conveniente saber algo de ecuaciones diferenciales ordinarias. Y para el capítulo 8 es aconsejable leerse antes el complemento B si no se han visto con anterioridad las ecuaciones diferenciales en derivadas parciales.

Entre los objetivos de este texto marcamos dos: su fácil comprensión para estudiantes universitarios con un conocimiento mínimo de matemáticas e instruir a los estudiantes para que practiquen con los métodos numéricos en un ordenador. Se ha intentado exponer los conceptos de una manera clara y sencilla, con muy pocos resultados teóricos formalmente enunciados y sin demostraciones.

En todos los capítulos se ha intentado ilustrar cada método numérico descrito con un ejemplo, que habitualmente ha sido desarrollado previamente en alguna de los textos que aparecen en la bibliografía, con el objetivo claro de animar a los estudiantes a tomar contacto con obras de referencia sobre la materia. También al final de cada capítulo se han añadido unas sugerencias bibliográficas y una colección de ejercicios propuestos con el objetivo de incitar a los estudiantes a que planteen problemas, comprendan los recursos teóricos necesarios y utilicen los métodos numéricos adecuados para resolver los problemas. Además se ha pretendido que el texto sirva de autoaprendizaje para los estudiantes.

Los temas tratados constituyen materia suficiente para un curso introductorio sobre métodos numéricos, y están organizados en torno a tres partes, cada una de las cuales está dividida en capítulos. Se empieza con temas sencillos y poco a poco se van complicando. En los cinco primeros capítulos se desarrollan las

técnicas más básicas de los métodos numéricos. Los dos siguientes están dedicados a la resolución numérica de ecuaciones diferenciales ordinarias. Y el último versa sobre la resolución numérica de ecuaciones diferenciales en derivadas parciales.

La primera parte del texto comienza con un capítulo introductorio en el que se hace especial hincapié en el papel tan importante que juegan los errores a la hora de implementar métodos numéricos en un ordenador para resolver problemas. Para ello se recuerdan los polinomios de Taylor y cómo se calculan y almacenan los números en un ordenador. El capítulo 2 trata sobre los métodos numéricos básicos para resolver sistemas de ecuaciones lineales, distinguiendo entre métodos directos y métodos iterativos. El capítulo 3 describe los métodos iterativos más conocidos para resolver ecuaciones no lineales, así como la resolución de sistemas no lineales mediante el método de Newton. En el capítulo 4 se estudia la interpolación polinómica de Lagrange, la interpolación mediante funciones *splines* y la aproximación de mínimos cuadrados. El capítulo 5 analiza la derivación y la integración numéricas a partir del polinomio de interpolación, describiendo diversas fórmulas, tanto para las derivadas como para las integrales.

La segunda parte del texto está dedicada a las ecuaciones diferenciales ordinarias. El capítulo 6 abarca los problemas de valor inicial, examinado los métodos de un paso y los métodos multipaso, tanto explícitos como implícitos, incluyendo métodos predictor-corrector, y terminando con la extensión a ecuaciones diferenciales de orden superior y sistemas de ecuaciones diferenciales. El capítulo 7 describe métodos numéricos para resolver problemas de valores en la frontera en dos puntos, discutiendo los métodos de disparo y los de diferencias finitas, distinguiendo los casos lineal y no lineal.

La tercera parte del texto, que comprende el capítulo 8, repasa los métodos numéricos para resolver ecuaciones diferenciales en derivadas parciales, describiendo métodos numéricos basados en la aproximación por diferencias finitas que son de dos tipos: explícitos e implícitos. También se hace una pequeña introducción al método de los elementos finitos.

Además de cubrir los temas estándares desarrollados en los capítulos 1–8, se han añadido dos complementos que los completan. El primero se dedica a la aproximación de valores y vectores propios mediante el método de la potencia. En el segundo se introducen las ecuaciones diferenciales en derivadas parciales, que habitualmente no se encuentra en textos de métodos numéricos, pero que aquí nos ha parecido interesante recordar debido a que a veces algunos estudiantes universitarios no las han visto con anterioridad.

Al final hemos adjuntado la bibliografía básica utilizada para elaborar este texto, junto con una bibliografía complementaria con la que los estudiantes puedan trabajar a la hora de profundizar en los métodos numéricos. También pueden encontrar por su cuenta múltiples referencias electrónicas en Internet. Una buena forma de comenzar es visitando las páginas de Wikipedia: `http://es.wikipedia.org/wiki/Analisis_numerico` (en español) y `http://en.wikipedia.org/wiki/Numerical_analysis` (en inglés).

Finalmente, quiero dar las gracias de forma especial a Mario Escario Gil, que durante los dos años que estuvo impartiendo docencia en la Universidad de La Rioja dió clases de métodos numéricos y elaboró la sección dedicada a la introducción del método de los elementos finitos. Para terminar quiero dejar constancia de mi agradecimiento al profesor Miguel Ángel Hernández Verón por inculcarme su interés por los métodos numéricos y mostrarse siempre dispuesto para la reflexión y el buen entendimiento de los mismos.

Logroño, enero de 2012 J. A. E. F

Contenidos

Capítulo 1

Preliminares matemáticos y computacionales

1.1. Introducción

En este texto se estudian problemas que se pueden resolver mediante métodos de aproximación, técnicas que se conocen genéricamente como **métodos numéricos**. Empezamos considerando algunos aspectos matemáticos y computacionales que aparecen cuando se aproxima la solución de un problema.

Dos son los objetivos de este capítulo. El primero es revisar algunos conceptos y términos fundamentales del cálculo y del álgebra matricial, que son útiles en la obtención de los métodos numéricos en sí mismos, y que sirvan como recordatorio de los conceptos con los que se supone que los estudiantes están familiarizados. Y el segundo es presentar algunos conceptos preparatorios importantes para el estudio de los métodos numéricos.

Como todos los métodos numéricos tienen la importante característica de que cometen errores, hay dos cosas que se deben tener en cuenta fundamentalmente a la hora de aplicar un método numérico para resolver un problema. La primera, y más obvia, es obtener la aproximación. La segunda, pero igualmente importante, es establecer la bondad de la aproximación; es decir, disponer de alguna medida, o por lo menos de una cierta idea, de su grado de exactitud. También es importante destacar una de las dificultades habituales que surgen cuando se usan técnicas para aproximar la solución de un problema: ¿dónde y por qué se producen errores en las operaciones aritméticas y cómo se pueden controlar? Es fundamental entonces identificar, cuantificar y minimizar dichos errores. Otro de los aspectos a tener en cuenta a la hora de aplicar un método numérico para resolver un problema es el coste operacional de los procesos numéricos.

Para las demostraciones y un mayor detalle pueden consultarse [7], [18] y [23].

1.2. Un recordatorio de cálculo

Comenzamos el capítulo con un repaso de algunos aspectos importantes del cálculo que son necesarios a lo largo del texto. Suponemos que los estudiantes que lean este texto conocen la terminología, la notación y los resultados que se dan en un curso típico de cálculo.

1.2.1. Límites y continuidad

El límite de una función en un punto dado dice, en esencia, a qué se aproximan los valores de la función cuando los puntos de su dominio se acercan a dicho punto dado, pero éste es un concepto difícil de establecer con precisión. La noción de límite es esencial para el cálculo infinitesimal.

• Decimos que una función $f(x)$ definida en un conjunto $\mathcal{S}$ de números reales tiene **límite** L en el punto $x = x_0$, lo que denotamos por

$$\lim_{x \to x_0} f(x) = L,$$

si dado cualquier número real $\varepsilon > 0$, existe un número real $\delta > 0$ tal que $|f(x) - L| < \varepsilon$ siempre que $0 < |x - x_0| < \delta$. (Esta definición asegura que los valores de la función estarán cerca de L siempre que x esté suficientemente cerca de x_0).

• Se dice que una función es continua en un punto de su dominio cuando el límite en dicho punto coincide con el valor de la función en él. Es decir, una función $f(x)$ es **continua** en el punto $x = x_0$ si

$$\lim_{x \to x_0} f(x) = f(x_0),$$

y se dice que f es **continua en el conjunto** $\mathcal{S}$ si es continua en cada uno de los puntos de $\mathcal{S}$. Denotaremos el conjunto de todas las funciones f que son continuas en $\mathcal{S}$ por $C(\mathcal{S})$. Cuando $\mathcal{S}$ sea un intervalo de la recta real, digamos $[a, b]$, entonces usaremos la notación $C[a, b]$.

• El límite de una sucesión de números reales o complejos se define de manera parecida. Decimos que una sucesión $\{x_n\}_{n=1}^{\infty}$ **converge** a un número x, lo que se escribe

$$\lim_{n \to \infty} x_n = x \quad (\text{o bien,} \quad x_n \to x \quad \text{cuando} \quad n \to \infty),$$

si, dado cualquier $\varepsilon > 0$, existe un número natural $N(\varepsilon)$ tal que $|x_n - x| < \varepsilon$ para cada $n > N(\varepsilon)$. Cuando una sucesión tiene límite, se dice que es una **sucesión convergente**.

• **Continuidad y convergencia de sucesiones.** Si $f(x)$ es una función definida en un conjunto $\mathcal{S}$ de números reales y $x_0 \in \mathcal{S}$, entonces las siguientes afirmaciones son equivalentes:

a. $f(x)$ es continua en $x = x_0$,

b. Si $\{x_n\}_{n=1}^{\infty}$ es cualquier sucesión en $\mathcal{S}$ que converge a x_0, entonces $\lim\limits_{n \to \infty} f(x_n) = f(x_0)$.

• **Teorema del valor intermedio o de Bolzano.** Si $f \in C[a, b]$ y ℓ es un número cualquiera entre $f(a)$ y $f(b)$, entonces existe al menos un número c en (a, b) tal que $f(c) = \ell$. Véase la figura 1.1.

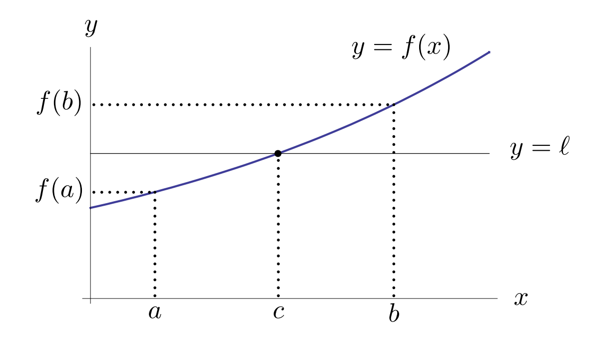


Figura 1.1: Teorema del valor intermedio o de Bolzano.

• Todas las funciones con las que vamos a trabajar en la discusión de los métodos numéricos serán continuas, ya que esto es lo mínimo que debemos exigir para asegurar que la conducta de un método se puede predecir.

1.2.2. Derivabilidad

• Si $f(x)$ es una función definida en un intervalo abierto que contiene un punto x_0, entonces se dice que $f(x)$ es **derivable** en $x = x_0$ cuando existe el límite

$$f'(x_0) = \lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0}.$$

El número $f'(x_0)$ se llama **derivada** de f en x_0 y coincide con la pendiente de la recta tangente a la gráfica de f en el punto $(x_0, f(x_0))$, tal y como se muestra en la figura 1.2.

Una función derivable en cada punto de un conjunto $\mathcal{S}$ se dice que es **derivable** en $\mathcal{S}$. La derivabilidad es una condición más fuerte que la continuidad en el siguiente sentido.

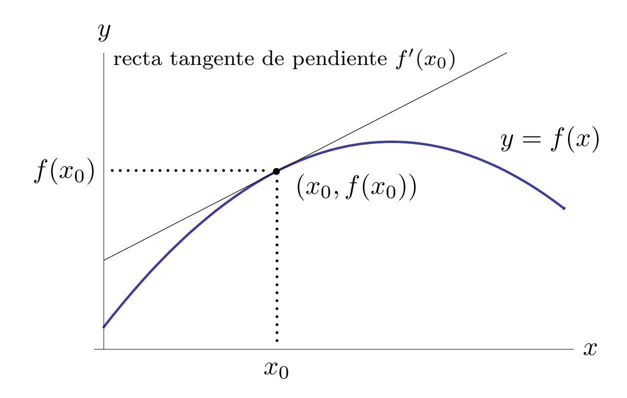


Figura 1.2: Derivada de una función en un punto.

• **Derivabilidad implica continuidad.** Si la función $f(x)$ es derivable en $x = x_0$, entonces $f(x)$ es continua en $x = x_0$.

• El conjunto de todas las funciones que admiten n derivadas continuas en $\mathcal{S}$ se denota por $C^n(\mathcal{S})$, mientras que el conjunto de todas las funciones indefinidamente derivables en $\mathcal{S}$ se denota por $C^\infty(\mathcal{S})$. Las funciones polinómicas, racionales, trigonométricas, exponenciales y logarítmicas están en $C^\infty(\mathcal{S})$, siendo $\mathcal{S}$ el conjunto de puntos en los que están definidas.

• **Teorema del valor medio o de Lagrange.** Si $f \in C[a, b]$ y es derivable en (a, b), entonces existe un punto c en (a, b) tal que

$$f'(c) = \frac{f(b) - f(a)}{b - a}.$$

Geométricamente hablando, véase la figura 1.3, el teorema del valor medio dice que hay al menos un número $c \in (a, b)$ tal que la pendiente de la recta tangente a la curva $y = f(x)$ en el punto $(c, f(c))$ es igual a la pendiente de la recta secante que pasa por los puntos $(a, f(a))$ y $(b, f(b))$.

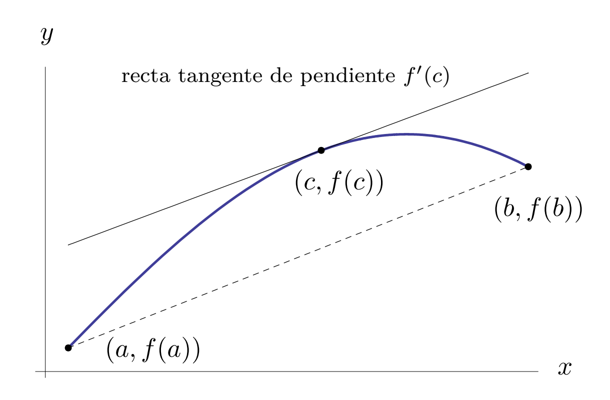


Figura 1.3: Teorema del valor medio o de Lagrange.

• **Teorema de los valores extremos.** [Este resultado se usa frecuentemente para establecer cotas del error cometido]. Si $f \in C[a, b]$, entonces existen c_1 y c_2 en (a, b) tales que $f(c_1) \leq f(x) \leq f(c_2)$ para todo x en $[a, b]$. Si además, f es derivable en (a, b), entonces los puntos c_1 y c_2 están en los extremos de $[a, b]$ o bien son puntos críticos, es decir, puntos en los que f' se anula.

1.2.3. Integración

• **Primer teorema fundamental o regla de Barrow.** Si $f \in C[a,b]$ y F es una primitiva cualquiera de f en $[a,b]$ (es decir, $F'(x) = f(x)$), entonces

$$\int_a^b f(x)\,dx = F(b) - F(a).$$

• **Segundo teorema fundamental.** Si $f \in C[a,b]$ y $x \in (a,b)$, entonces

$$\frac{d}{dx}\int_a^x f(t)\,dt = f(x).$$

• **Teorema del valor medio para integrales.** Si $f \in C[a,b]$, g es integrable en $[a,b]$ y $g(x)$ no cambia de signo en $[a,b]$, entonces existe un punto c en (a,b) tal que

$$\int_a^b f(x)g(x)\,dx = f(c)\int_a^b g(x)\,dx.$$

Cuando $g(x) \equiv 1$, véase la figura 1.4, este resultado es el habitual teorema del valor medio para integrales y proporciona el valor medio de la función f en el intervalo $[a,b]$, que está dado por

$$f(c) = \frac{1}{b-a}\int_a^b f(x)\,dx.$$

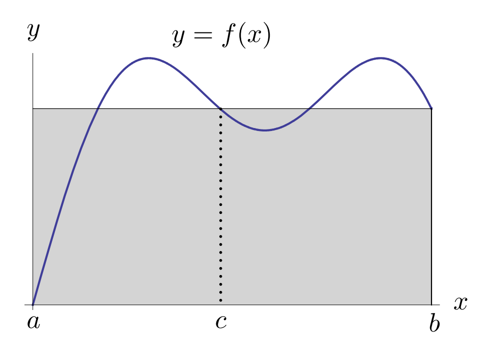


Figura 1.4: Teorema del valor medio para integrales.

1.2.4. Polinomios de Taylor

Terminamos este repaso al cálculo con los polinomios de Taylor. Siempre nos quedaremos cortos al hacer hincapié sobre la importancia de los polinomios de Taylor en el análisis numérico; en particular, el siguiente resultado se usa una y otra vez.

• **Teorema de Taylor.** Supongamos que $f \in C^n[a,b]$ y que $f^{(n+1)}$ existe en $[a,b]$. Sea x_0 un punto en $[a,b]$. Entonces, para cada x en $[a,b]$, existe un punto $\xi(x)$ entre x_0 y x tal que

$$f(x) = P_n(x) + R_n(x),$$

donde

$$P_n(x) = f(x_0) + f'(x_0)h + \frac{f''(x_0)}{2!}h^2 + \cdots + \frac{f^{(n)}(x_0)}{n!}h^n = \sum_{k=0}^{n}\frac{f^{(k)}(x_0)}{k!}h^k,$$

$$R_n(x) = \frac{f^{(n+1)}(\xi(x))}{(n+1)!}h^{n+1} \qquad \text{y} \qquad h = x - x_0.$$

El polinomio $P_n(x)$ se llama **n-ésimo polinomio de Taylor** de f alrededor de x_0 (véase la figura 1.5) y $R_n(x)$ se llama **error de truncamiento** (o *resto de Taylor*) asociado a $P_n(x)$. Como el punto $\xi(x)$ en el error de truncamiento $R_n(x)$ depende del punto x en el que se evalúa el polinomio $P_n(x)$, podemos verlo como una función de la variable x. Sin embargo, no debemos confiar en que seremos capaces de determinar explícitamente la función $\xi(x)$; el Teorema de Taylor simplemente asegura que dicha función existe y que sus valores están entre x y x_0. De hecho, uno de los problemas habituales de los métodos numéricos es la determinación de una cota realista del valor $f^{(n+1)}(\xi(x))$ para los puntos x de un cierto intervalo dado.

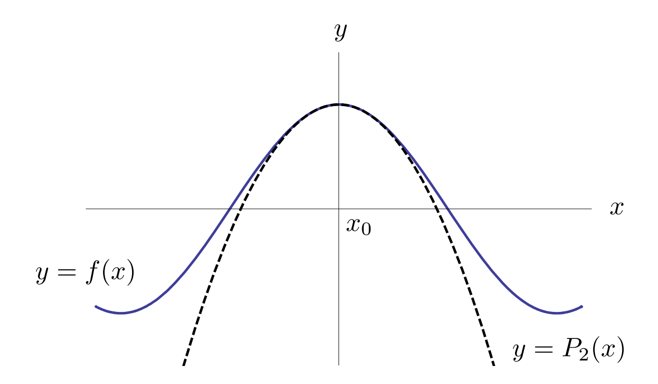


Figura 1.5: Gráficas de $y = f(x)$ (línea continua) y de su polinomio de Taylor $y = P_2(x)$ alrededor de x_0 (línea discontinua).

• La serie infinita que resulta al tomar límite en la expresión de $P_n(x)$ cuado $n \to \infty$ se llama *serie de Taylor* de f alrededor de x_0. Cuando $x_0 = 0$, el polinomio de Taylor se suele denominar **polinomio de Maclaurin**, y la serie de Taylor se llama *serie de Maclaurin*.

• La denominación *error de truncamiento* en el teorema de Taylor se refiere al error que se comete al usar una suma truncada (es decir, finita) al aproximar la suma de una serie infinita.

• **Teorema de Taylor en dos variables.** Si $f(x,y)$ y todas sus derivadas parciales de orden menor o igual que $n+1$ son continuas en $\mathcal{D} = \{(x,y)|\ a \le x \le b,\ c \le y \le d\}$ y los puntos (x,y) y $(x+h, y+k)$ están ambos en $\mathcal{D}$, entonces, para cada x en $[a,b]$, existe un punto $\xi(x)$ entre x y $x+h$, y, para cada y en $[c,d]$, existe un punto $\nu(x)$ entre y e $y+k$, tales que

$$f(x,y) = P_n(x,y) + R_n(x,y),$$

donde

$$\begin{aligned} P_n(x,y) &= f(x,y) + \left(h\frac{\partial f}{\partial x}(x,y) + k\frac{\partial f}{\partial y}(x,y) \right) \\ &\quad + \left(\frac{h^2}{2}\frac{\partial^2 f}{\partial x^2}(x,y) + hk\frac{\partial^2 f}{\partial x\,\partial y}(x,y) + \frac{k^2}{2}\frac{\partial^2 f}{\partial y^2}(x,y) \right) + \cdots + \frac{1}{n!}\sum_{j=0}^{n}\binom{n}{j}h^{n-j}k^j\frac{\partial^n f}{\partial x^{n-j}\,\partial y^j}(x,y), \\ R_n(x,y) &= \frac{1}{(n+1)!}\sum_{j=0}^{n+1}\binom{n+1}{j}h^{n+1-j}k^j\frac{\partial^{n+1} f}{\partial x^{n+1-j}\partial y^j}(\xi(x),\nu(x)). \end{aligned}$$

1.3. Un recordatorio de álgebra matricial

Las matrices se utilizan ampliamente en computación debido a su facilidad y ligereza a la hora de manipular información, así que son muy utilizadas en cálculo numérico con ordenador.

1.3.1. Matrices

• Una **matriz** es un conjunto bidimensional de escalares, llamados *elementos*, ordenados en filas y columnas en forma de tabla rectangular. Indicamos el tamaño (o *dimensión*) de la matriz con el número de filas y columnas que contiene. Una matriz de m filas y n columnas, o matriz (de orden) $m \times n$, es un conjunto de $m \cdot n$ elementos a_{ij}, con $i = 1, 2, \dots, m$ y $j = 1, 2, \dots, n$, que se representa de la siguiente forma:

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}.$$

Podemos abreviar la representación de la matriz anterior de la forma $A = (a_{ij})$ con $i = 1, 2, \dots, m$ y $j = 1, 2, \dots, n$.

• Hay una relación directa entre matrices y vectores puesto que podemos pensar una matriz como una composición de vectores filas o de vectores columnas. Además, un vector es un caso especial de matriz: un **vector fila** es una matriz con una sola fila y varias columnas, y un **vector columna** es una matriz con varias filas y una sola columna. En el caso $m = n = 1$, la matriz designa simplemente un escalar.

1.3.2. Operaciones con matrices

• Si $A = (a_{ij})$ y $B = (b_{ij})$ son dos matrices que tienen el mismo orden, $m \times n$, decimos que A y B son **iguales** si $a_{ij} = b_{ij}$ para todo $i = 1, 2, \dots, m$ y $j = 1, 2, \dots, n$.

• Si $A = (a_{ij})$ y $B = (b_{ij})$ son dos matrices que tienen el mismo orden $m \times n$, la **suma** de A y B es una matriz $C = (c_{ij})$ del mismo orden $m \times n$ con $c_{ij} = a_{ij} + b_{ij}$ para todo $i = 1, 2, \dots, m$ y $j = 1, 2, \dots, n$.

• Si $A = (a_{ij})$ es una matriz de orden $m \times n$, la **multiplicación** de A **por un escalar** λ, es una matriz $C = (c_{ij})$ del mismo orden $m \times n$ con $c_{ij} = \lambda\, a_{ij}$ para todo $i = 1, 2, \dots, m$ y $j = 1, 2, \dots, n$.

• Si $A = (a_{ij})$ es una matriz de orden $m \times n$, la matriz **traspuesta** de A es la matriz que resulta de intercambiar sus filas por sus columnas, se denota por A^T y es de orden $n \times m$.

• Si $A = (a_{ij})$ es una matriz de orden $m \times p$ y $B = (b_{ij})$ es una matriz de orden $p \times n$, el **producto** de A por B es una matriz $C = (c_{ij})$ de orden $m \times n$ con $c_{ij} = \sum_{k=1}^{p} a_{ik} b_{kj}$ para todo $i = 1, 2, \dots, m$ y $j = 1, 2, \dots, n$. Obsérvese que el producto de dos matrices solo está definido si el número de columnas de la primera matriz coincide con el número de filas de la segunda.

1.3.3. Matrices especiales

• Decimos que una matriz $A = (a_{ij})$ es **cuadrada** si tiene el mismo número de filas que de columnas, y de orden n si tiene n filas y n columnas. Se llama *diagonal principal* al conjunto de elementos $a_{11}, a_{22}, \dots, a_{nn}$.

• Llamamos matriz **diagonal** a una matriz cuadrada que tiene algún elemento distinto de cero en la diagonal principal y ceros en el resto de elementos.

• Una matriz cuadrada con ceros en todos los elementos por encima (debajo) de la diagonal principal se llama matriz **triangular inferior** (**superior**).

• Una matriz diagonal con todo unos en la diagonal principal se denomina matriz **identidad** y se denota por I. Es por definición la única matriz cuadrada tal que $AI = IA = A$ para cualquier matriz cuadrada A.

• Una matriz **simétrica** es una matriz cuadrada A tal que $A = A^T$.

• La matriz **cero** es una matriz con todos sus elementos iguales a cero.

1.3.4. Inversa de una matriz

• Decimos que una matriz cuadrada A es *invertible* (o *regular* o *no singular*) si existe una matriz cuadrada B tal que $AB = BA = I$. Se dice entonces que B es la matriz **inversa** de A y se denota por A^{-1}. (Una matriz que no es invertible se dice *singular*.)

• Si una matriz A es invertible, su inversa también lo es y $\left(A^{-1}\right)^{-1} = A$.

• Si A y B son dos matrices invertibles, su producto también lo es y $(AB)^{-1} = B^{-1}A^{-1}$.

1.3.5. Determinante de una matriz

• El **determinante** de una matriz solo está definido para matrices cuadradas y su valor es un escalar. El determinante de una matriz A cuadrada de orden n se denota por $|A|$ o $\det(A)$ y se define como

$$\det(A) = \sum_{j} (-1)^k a_{1j_1} a_{2j_2} \cdots a_{nj_n},$$

donde la suma se toma para todas las $n!$ permutaciones de grado n y k es el número de intercambios necesarios para poner el segundo subíndice en el orden $1, 2, \dots, n$.

• Algunas propiedades de los determinantes son:

a. $\det(A) = \det(A^T)$.

b. $\det(AB) = \det(A)\det(B)$.

c. $\det(A^{-1}) = \frac{1}{\det(A)}$.

d. Si dos filas o dos columnas de una matriz coinciden, el determinante de esta matriz es cero.

e. Cuando se intercambian dos filas o dos columnas de una matriz, su determinante cambia de signo.

f. El determinante de una matriz diagonal es el producto de los elementos de la diagonal.

• Si denotamos por A_{ij} la matriz de orden $(n-1)$ que se obtiene de eliminar la fila i y la columna j de la matriz A, llamamos **menor complementario** asociado al elemento a_{ij} de la matriz A al $\det(A_{ij})$.

• Se llama k-ésimo **menor principal** de la matriz A al determinante de la submatriz principal de orden k.

• Definimos el **cofactor** del elemento a_{ij} de la matriz A por $\Delta_{ij} = (-1)^{i+j}\det(A_{ij})$.

• Si A es una matriz invertible de orden n, entonces $A^{-1} = \frac{1}{\det(A)} C$, donde C es la matriz de elementos Δ_{ij}, para todo $i, j = 1, 2, \dots, n$. Obsérvese entonces que una matriz cuadrada es invertible si y solo si su determinante es distinto de cero.

1.3.6. Valores propios y vectores propios

• Si A es una matriz cuadrada de orden n, un número λ es un **valor propio** de A si existe un vector no nulo $\mathbf{v}$ tal que $A\mathbf{v} = \lambda\mathbf{v}$. Al vector $\mathbf{v}$ se le llama **vector propio** asociado al valor propio λ.

• El valor propio λ es solución de la **ecuación característica** $\det(A - \lambda I) = 0$, donde $\det(A - \lambda I)$ se llama **polinomio característico**. Este polinomio es de grado n en λ y tiene n valores propios (no necesariamente distintos).

1.3.7. Normas vectoriales y normas matriciales

• Para medir la «longitud» de los vectores y el «tamaño» de las matrices se suele utilizar el concepto de **norma**, que es una función que toma valores reales. Un ejemplo simple en el espacio euclidiano tridimensinal es un vector $\mathbf{v} = (v_1, v_2, v_3)$, donde v_1, v_2 y v_3 son las distancias a lo largo de los ejes x, y y z, respectivamente. La longitud del vector $\mathbf{v}$ (es decir, la distancia del punto $(0, 0, 0)$ al punto (v_1, v_2, v_3)) se calcula como $\|\mathbf{v}\|_e = \sqrt{v_1^2 + v_2^2 + v_3^2}$, donde la notación $\|\mathbf{v}\|_e$ indica que esta longitud se refiere a la *norma euclidiana* del vector $\mathbf{v}$. De forma similar, para un vector $\mathbf{v}$ de dimensión n, $\mathbf{v} = (v_1, v_2, \ldots, v_n)$, la norma euclidiana se calcula como $\|\mathbf{v}\|_e = \sqrt{v_1^2 + v_2^2 + \cdots + v_n^2}$. Este concepto puede extenderse a una matriz $m \times n$, $A = (a_{ij})$, de la siguiente manera $\|A\|_e = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}^2}$, que recibe el nombre de *norma de Frobenius*.

• Hay otras alternativas a las normas euclidiana y de Frobenius. Dos normas usuales son la *norma 1* y la *norma infinito*:

a. La norma 1 de un vector $\mathbf{v} = (v_1, v_2, \ldots, v_n)$ se define como $\|\mathbf{v}\|_1 = \sum_{i=1}^{n} |v_i|$. De forma similar, la norma 1 de una matriz $m \times n$, $A = (a_{ij})$, se define como $\|A\|_1 = \max_{1 \leq j \leq n} \sum_{i=1}^{m} |a_{ij}|$.

b. La norma infinito de un vector $\mathbf{v} = (v_1, v_2, \ldots, v_n)$ se define como $\|\mathbf{v}\|_\infty = \max_{i \leq 1 \leq n} |v_i|$. Similarmente, la norma infinito de una matriz $m \times n$, $A = (a_{ij})$, se define como $\|A\|_\infty = \max_{1 \leq i \leq m} \sum_{j=1}^{n} |a_{ij}|$.

• Recordemos, por la teoría del álgebra lineal, que todas las normas son equivalentes en un espacio vectorial de dimensión finita. Por ejemplo, si dos vectores están próximos según la norma $\| \cdot \|_1$, también lo están según la norma $\| \cdot \|_e$, y recíprocamente.

• Aunque puede haber ciertas ventajas teóricas para la utilización de algunas normas, la elección de una norma suele estar influenciada por las consideraciones prácticas. Por ejemplo, la norma infinita es muy utilizada por la facilidad con que se calcula y por el hecho de que habitualmente proporciona una medida adecuada del tamaño de una matriz.

1.4. Algunas ideas básicas sobre el cálculo con ordenador

La utilización de ordenadores ha cambiado la forma de resolver problemas numéricamente. A la hora de hacer cálculos con lápiz y papel solemos utilizar aritmética racional, mientras que si esos mismos cálculos los hacemos con el ordenador, en la mayoría de los casos, podemos utilizar aritmética finita y guardar los valores obtenidos en la memoria del ordenador, lo que habitualmente conlleva una pérdida de exactitud. Por lo tanto, al realizar cálculos con el ordenador se cometen errores y el análisis de estos errores es importante.

En esta sección se describen aspectos básicos relacionados con la identificación, cuantificación y minimización de errores. Destacaremos que dos son los errores numéricos más comunes: los errores de redondeo y los errores de trucamiento. Los primeros son como consecuencia de que los ordenadores solo pueden representar cantidades con un número finito de dígitos. Los segundos vienen como consecuencia de la diferencia entre una formulación matemática y su aproximación obtenida mediante un método numérico.

Como es normal que los errores se propaguen a lo largo de una sucesión de operaciones, habrá que prestar especial atención al fenómeno de la propagación de los errores, destacando los métodos que proporcionen resultados fiables. Además, como bastantes métodos numéricos están definidos mediante técnicas iterativas que involucran sucesiones numéricas, daremos una definición para describir la rapidez de su convergencia, estando en general interesados en que ésta sea lo más rápida posible. Terminaremos hablando del coste computacional de los métodos numéricos.

1.4.1. Fuentes de error

Comenzamos repasando algunos conceptos básicos referentes a la representación aproximada de los números. A la hora de realizar un cálculo es importante asegurarse de que los números que intervienen en el cálculo se pueden utilizar con confianza. Para ello, se introduce el concepto de **cifras significativas**, que designan formalmente la confianza de un valor numérico, y que son aquellas que pueden utilizarse con confianza.

Otros conceptos a los que también hay que prestar cierta atención son la **exactitud** y la **precisión**. La exactitud mide la cercanía entre los valores calculados y los valores exactos, mientras que la precisión se refiere a la cercanía entre distintos valores calculados.

Definiciones de error

• A la hora de representar cantidades y cálculos matemáticos exactos mediante aproximaciones aparecen errores numéricos, que incluyen los errores de redondeo que se producen cuando se representan números exactos mediante números con límite de cifras significativas, y los errores de truncamiento, que se produden al emplear aproximaciones como procedimientos matemáticos exactos. En ambos casos el error numérico viene dado por la diferencia entre el valor exacto y su valor aproximado, lo que se denomina **error absoluto**. Esto es, si $\tilde{x}$ es el valor aproximado y x el valor exacto (habitualmente desconocido), el error cometido a la hora de utilizar el valor aproximado es: $Error\ absoluto\,(x) = x - \tilde{x}$. En la mayoría de las aplicaciones el signo del error absoluto es de poca importancia, así que habitualmente se considera el error absoluto sin signo: $Error\ absoluto\,(x) = |x - \tilde{x}|$.

Sin embargo, para problemas en los que la magnitud del valor real puede ser particularmente muy grande (o muy pequeña), el **error relativo** puede ser más importante que el error absoluto ($Error\ relativo\,(x) = |x - \tilde{x}|/|x|$). Cuando queremos calcular el error relativo, a menudo se utiliza el valor aproximado $\tilde{x}$ en el denominador en vez del valor exacto desconocido x.

En la práctica es imposible conocer exactamente el valor del error correspondiente a una aproximación, de manera que deberemos contentarnos con acotarlo o estimarlo (sin conocerlo). Muchos métodos numéricos proporcionan una estimación del error, además de la aproximación, esperando que dicha estimación coincida aproximadamente con el error.

• En general, es obvio que los programas de ordenador solo pueden guardar un número finito de cifras significativas, de manera que cantidades específicas como $2/3$, $\sqrt{3}$, e o π no se pueden representar con exactitud con un ordenador, puesto que tienen infinitos decimales. Existen dos posibilidades. La más simple es truncar el número, descartando todas las cifras que estén detrás de las que que el ordenador puede guardar. La otra es redondear el número, el resultado depende entonces del valor de la primera cifra a descartar. Si el ordenador permite hasta n cifras, el redondeo produce el mismo resultado que el truncamiento si la cifra $(n+1)$-ésima es 0, 1, 2, 3 o 4, mientras que si la cifra $(n+1)$-ésima es 5, 6, 7, 8 o 9, entonces la cifra n-ésima se incremente en 1. Cuando se omiten cifras significativas hablaremos de **error de redondeo**.

Los errores que se cometen con el redondeo tienen menor probabilidad de acumularse durante la repetición de cálculos, puesto que el valor exacto es más grande que el valor redondeado aproximadamente la mitad de las veces y más pequeño aproximadamente la otra mitad. Además, el error absoluto más grande que puede ocurrir es unas dos veces más grande tanto a la hora de truncar como a la hora de redondear. Por otra parte, truncar no requiere decidir si hay que cambiar la última cifra guardada.

Los errores de redondeo se producen frecuentemente cuando los números que están implicados en los cálculos difieren significativamente en su magnitud y cuando se restan dos números que son casi idénticos.

• Los **errores de truncamiento** se producen cuando utilizamos una aproximación en lugar de un procedimiento matemático exacto. Para conocer las características de estos errores se suelen considerar los polinomios de Taylor, que se utilizan ampliamente en los métodos numéricos para expresar funciones de manera aproximada. Una gran cantidad de métodos numéricos están basados en la utilización de unos pocos términos de los polinomios de Taylor para aproximar una función.

Cuando se aproxima un proceso continuo por uno discreto, para errores provocados por un tamaño de paso finito h, resulta a menudo útil describir la dependencia del error de h cuando h tiende a cero.

> Decimos que una función $f(h)$ es una «O grande» de h^n si $|f(h)| \leq c|h^n|$ para alguna constante c, cuando h es próximo a cero; se escribe $f(h) = \mathcal{O}(h^n)$.

Si un método tiene un término de error que es $\mathcal{O}(h^n)$, se suele decir que es un **método de orden** n. Por ejemplo, si utilizamos un polinomio de Taylor para aproximar la función g en $x = a + h$, tenemos

$$g(x) = g(a+h) = g(a) + hg'(a) + \frac{h^2}{2!}g''(a) + \frac{h^3}{3!}g'''(\xi), \quad \text{para algún} \quad \xi \in [a, a+h].$$

Suponiendo que g es suficientemente derivable, la aproximación anterior es $\mathcal{O}(h^3)$, puesto que el error, $\frac{h^3}{3!}g'''(\xi)$, satisface

$$\left|\frac{h^3}{3!}\,g'''(\xi)\right| \leq \frac{M}{3!}\left|h^3\right|,$$

donde M es el máximo de $g'''(x)$ para $x \in [a, a+h]$.

El error de truncamiento depende del método numérico utilizado para resolver un problema, es independiente del error de redondeo y se produce incluso cuando las operaciones matemáticas son exactas.

• La suma de los errores de redondeo y de truncamiento es lo que habitualmente llamamos **error numérico total** (también llamado error verdadero), que está incluido en la solución numérica. En general, si se incrementa el número de cifras significativas en el ordenador, se minimizan los errores de redondeo, y los errores de truncamiento disminuyen a medida que los errores de redondeo se incrementan. Por lo tanto, para disminuir uno de los dos sumandos del error total debemos incrementar el otro. El reto consiste en identificar dónde los errores de redondeo no muestren los beneficios de la reducción del error de truncamiento y determinar el tamaño del incremento apropiado para un cálculo en particular. Estas situaciones son poco comunes en casos reales porque habitualmente los ordenadores utilizan suficientes cifras significativas como para que los errores de redondeo no predominen.

Como el error total no se puede calcular en la mayoría de los casos, se suelen utilizar otras medidas para estimar la exactitud de un método numérico, que suelen depender del método específico. En algunos métodos el error numérico se puede acotar, mientras que en otros se determina una estimación del orden de magnitud del error.

• Terminamos recordando que cuando buscamos las soluciones numéricas de un problema real los resultados que obtenemos generalmente no son exactos. Una fuente habitual de inexactitudes radica en la simplificación del modelo del problema original. También suelen aparecer errores a la hora de interpretar una colección de datos. Además, habrá que estar atentos a los errores que no están relacionados directamente con los métodos numéricos aplicados, como por ejemplo, entre otros, equivocaciones, errores de formulación o del modelo, e incertidumbre en la obtención de datos.

1.4.2. Representación de los números reales en el ordenador

La utilización de ordenadores hoy en día juega un papel fundamental en el desarrollo de los métodos numéricos, ya que nos permiten resolver problemas que no seríamos capaces de hacerlo sin ellos. Sin embargo, aunque los ordenadores están programados para que se puedan utilizar con una gran exactitud, hay que tener en cuenta algunas consecuencias de cómo realizan los cálculos, como consecuencia de que pueden conducir a efectos inesperados en los resultados.

Toda operación que realiza un ordenador («operación máquina») está sujeta a errores de redondeo, que aparecen como consecuencia de que en un ordenador no se puede representar más que un subconjunto finito del conjunto de los números reales.

Mientras que el conjunto de los números reales $\mathbb{R}$ es conocido, la manera en la que los ordenadores los tratan es menos conocida. Por una parte, como los ordenadores tienen recursos limitados, sólo se puede representar un subconjunto de $\mathbb{R}$ de dimensión finita. Denotamos, tal y como se hace en [23], este subconjunto por $\mathbb{F}$, que está formado por números que se llaman *números de punto flotante*. Por otra parte, como veremos posteriormente, $\mathbb{F}$ está caracterizado por propiedades diferentes de las de $\mathbb{R}$. La razón es que cualquier número real x es truncado en principio por el ordenador, dando origen a un nuevo número, llamado *número de punto flotante*, que denotamos por $fl(x)$ y que no coincide necesariamente con el número original x.

A finales de los años 1950 se estandarizó el uso de números de punto flotante para los ordenadores, aunque ciertas especificaciones todavía podían variar de un ordenador a otro, hasta que el Instituto de Ingenieros Electrónicos y Eléctricos (IEEE, en sus siglas en inglés) desarrolló el estándar IEEE en 1985, que fue aprobado en 1989 por la Comisión Internacional Electrónica (IEC, en sus siglas en inglés) como estándar internacional IEC559. La idea básica de la aritmética de punto flotante es esencialmente la misma que para la notación científica, excepto que para un ordenador la base es casi siempre 2 en vez de 10.

Representación en punto flotante

Un ordenador almacena, en general, un número real x de la siguiente forma ([23]):

$$x = (-1)^s \cdot (0.a_1 a_2 \dots a_t) \cdot \beta^e = (-1)^s \cdot m \cdot \beta^{e-t}, \quad a_1 \neq 0, \tag{1.1}$$

donde s es 0 o 1, β (un entero positivo mayor o igual que 2) es la *base* adoptada por el ordenador específico que estemos manejando, m es un entero llamado *mantisa* cuya longitud t es el máximo número de cifras a_i

(con $0 \le a_i \le \beta - 1$) que se almacenan, y e es un número entero llamado *exponente*. Los números de la forma (1.1) se llaman **números de punto flotante**, porque la posición de su punto decimal no es fija. A las cifras $a_1 a_2 \ldots a_p$ (con $p \le t$) se les suele llamar p primeras *cifras significativas* de x.

La condición $a_1 \ne 0$ asegura que un número no puede tener múltiples representaciones. Por ejemplo, sin esta restricción el número $1/10$ podría ser representado (en la base decimal) como $0.1 \cdot 10^0$, pero también como $0.01 \cdot 10^1$, etc.

Por tanto, el conjunto $\mathbb{F}$ está totalmente caracterizado por la base β, el número de cifras significativas t y el rango (L, U) (con $L < 0$ y $U > 0$) de variación del índice e. Por esto se denota como $\mathbb{F}(\beta, t, L, U)$.

El número 0 no pertenece a $\mathbb{F}$, pues en tal caso tendríamos $a_1 = 0$ en (1.1); por tanto se maneja separadamente. Además, como L y U son finitos, no se pueden representar números cuyo valor absoluto sea arbitrariamente grande o arbitrariamente pequeño. Siendo más concretos, el número real positivo más pequeño y el más grande de $\mathbb{F}$ vienen dados respectivamente por

$$x_{min} = \beta^{L-1} \qquad \text{y} \qquad x_{max} = \beta^U (1 - \beta^{-t}).$$

Es inmediato verificar que si $x \in \mathbb{F}(\beta, t, L, U)$, entonces $-x \in \mathbb{F}(\beta, t, L, U)$, de manera que $x_{min} \le |x| \le x_{max}$. Por lo tanto, solo podemos representar con el ordenador un número máximo en valor absoluto de números reales, de manera que cuando intentamos utilizar un número de valor absoluto mayor que el máximo representable, se produce un error llamado *overflow* (desbordamiento por exceso). También puede ocurrir lo contrario, cuando utilizamos un número no nulo pero menor en valor absoluto, produciéndose entonces un error llamado *underflow* (desbordamiento por defecto).

NOTA. ([23]) Si bien es cierto que los errores de redondeo son normalmente pequeños, cuando se repiten dentro de largos y complejos algoritmos, pueden dar origen a efectos catastróficos. Dos casos destacados conciernen a la explosión del cohete Arianne el 4 de junio de 1996, generada por un *overflow* en el ordenador de a bordo, y al fracaso de la misión de un misil americano *patriot* durante la guerra del Golfo en 1991, a causa de un error de redondeo en el cálculo de su trayectoria.

Épsilon de la máquina

Además de la importancia de la representación de los números reales en el ordenador, también es importante la exactitud con que se realizan los cálculos con los números reales. Para esto habrá que tener en cuenta el mayor valor positivo ϵ_M tal que $1 + x = 1$, para todo $x \in (0, \epsilon_M)$, que proporciona la distancia entre 1 y el número en punto flotante mayor que 1 más cercano a éste que se puede representar con el ordenador. Este número ϵ_M se llama *épsilon de la máquina* y caracteriza la precisión de la aritmética en punto flotante, dependiendo del ordenador y del tipo de variable. Para el estándar IEC559, $\epsilon_M = \beta^{1-t}$. En general, el valor ϵ_M depende del tipo de redondeo que utilice el sistema.

Por otra parte, notemos que el *error de redondeo* que se genera inevitablemente siempre que un número real $x \ne 0$ se reemplaza por su represetante $fl(x)$ en $\mathbb{F}$, es afortunadamente pequeño, puesto que

$$\left| \frac{x - fl(x)}{x} \right| \le \frac{1}{2} \epsilon_M. \tag{1.2}$$

Señalamos que en (1.2) estimamos el *error relativo* sobre x, que es indudablemente más significativo que el *error absoluto* $|x - fl(x)|$. En realidad, este último no tiene en cuenta el orden de magnitud de x mientras que el primero si.

Sobre las operaciones en punto flotante

Puesto que $\mathbb{F}$ es un subconjunto de $\mathbb{R}$, las operaciones algebraicas elementales sobre números de punto flotante no gozan de todas las propiedades de las operaciones análogas en $\mathbb{R}$. Concretamente, la conmutatividad para la suma se verifica (esto es $fl(x + y) = fl(y + x)$), así como para la multiplicación ($fl(xy) = fl(yx)$), pero se violan otras propiedades como la asociativa y la distributiva. Además, el 0 ya no es único. Véase [22] para mayor detalle.

El estándar IEC559 tiene en cuenta dos circunstancias excepcionales: el resultado de dividir un número finito por cero, que se denota por `Inf`, y el resultado de una operación indeterminada, como por ejemplo $0/0$ o ∞/∞, que produce lo que se llama un «no-número», denotado por `NaN` (del inglés *Not a Number*), al que no se aplican las reglas normales de cálculo (la presencia de un `NaN` en una sucesión de operaciones implica que automaticamente el resultado es un `NaN`). El tratamiento de ambos depende del entorno computacional, debiéndose evitar siempre que se pueda.

Almacenamiento de un número en la memoria del ordendor

Los ordenadores guardan y procesan los números en forma binaria (base 2). Cada dígito binario (0 o 1) se llama *bit* (por dígito binario). La memoria de los ordenadores está organizada en *bytes*, siendo un byte ocho bits. Según el estándar IEC559, los ordenadores guardan los números y realizan los cálculos en *precisión simple* o en *doble precisión*. Esto es que los números se guardan en cadenas de 32 bits en precisión simple y de 64 bits en doble precisión. En ambos casos el primer bit es para el signo del número, los siguientes 8 bits en precisión simple (11 bits en doble precisión) son para guardar el exponente y los siguientes 23 bits en precisión simple (52 bits en doble precisión) son para guardar la mantisa.

La *precisión* se refiere aquí al número de cifras significativas de un número real que se puede guardar en un ordenador. Por ejemplo, el número $1/9 = 0.1111\ldots$ solo se puede representar en un ordenador de forma truncada o redondeada con un número finito de dígitos binarios, ya que la cantidad de memoria donde se guardan estos bits es finita. Cuantas más cifras a la derecha del punto decimal se puedan guardar, más *precisa* es la representación del número en el ordenador. Notemos que la precisión en un número real en doble precisión no se dobla si comparamos con el número en precisión simple, sino que se refiere a que la doble precisión utiliza el doble de dígitos binarios para representar el número real que la precisión simple.

La doble precisión tiene la ventaja de poder representar más números de manera más exacta, mientras que la precisión simple tiene la ventaja de que requiere menos espacio, lo que puede ser importante cuando se guardan grandes cantidades de datos. En la práctica, como las velocidades de los ordenadores de hoy en día son más o menos las mismas para los dos formatos, se suele utilizar más la doble precisión.

En el estándar IEC559 para la aritmética binaria de punto flotante tenemos $\mathbb{F} = \mathbb{F}(2, 24, -125, 128)$ para la precisión simple y $\mathbb{F} = \mathbb{F}(2, 53, -1024, 1024)$ para la doble precisión. Notemos que, en doble precisión, 53 cifras significativas en base 2 corresponden a 15 cifras significativas en base 10.

A modo de resumen, podemos decir que la precisión del ordenador es la que dicta la exactitud de los resultados.

1.4.3. Estabilidad y convergencia

En realidad podemos decir que el error es inevitable en el cálculo numérico. Como acabamos de ver, el simple hecho de utilizar un ordenador para representar los números reales introduce errores. Por tanto, en parte, lo importante no es esforzarse por eliminar los errores, sino más bien ser capaces de controlar sus efectos.

Las pérdidas de exactitud debidas a los errores de redondeo se pueden evitar a menudo eligiendo con cuidado el orden en el que se realizan las operaciones o reformulando el problema.

A lo largo de este texto consideraremos varios problemas de aproximación y se necesitará, en cada caso, desarrollar métodos de aproximación que produzcan resultados precisos para una amplia variedad de problemas. Como los métodos de aproximación no se construyen de la misma forma, hace falta disponer de una serie de condiciones que nos permitan categorizar su precisión (aunque puede ser que no todas estas condiciones sean aplicables en cada problema particular).

Una condición que hay que imponer siempre que sea posible es la de **estabilidad**. Se dice que un método es *estable* si pequeñas variaciones en los datos iniciales producen pequeñas variaciones en los resultados finales correspondientes. Cuando pequeñas variaciones en los datos iniciales producen variaciones grandes en los resultados finales el método es *inestable*. Algunos métodos son estables sólo para algunos datos iniciales, se dice entonces que estos métodos son *condicionalmente inestables*. Es importante caracterizar las propiedades de estabilidad siempre que se pueda.

La forma en que los errores de redondeo crecen conforme se va aplicando un método es uno de los aspectos más importantes en relación con la estabilidad. Un método es estable si presenta crecimiento lineal del error (esto es, si existe una constante C independiente de n tal que $E_n \approx CnE_0$, donde $E_0 > 0$ denota un error inicial y E_n representa la magnitud del error después de n operaciones posteriores), mientras que es inestable si el crecimiento del error es exponencial (es decir, si existe una constante $C > 1$ independiente de n tal que $E_n \approx C^n E_0$).

Por otra parte, como a menudo se utilizan técnicas que construyen una sucesión de aproximaciones, lo que generalmente interesa es disponer de técnicas cuyas sucesiones converjan lo más rápidamente posible. Mediante la siguiente definición podremos comparar las velocidades de convergencia de los distintos métodos.

Supongamos que $\{\alpha_n\}_{n=1}^{\infty}$ es una sucesión que converge a un número α cuando $n \to \infty$. Si existen

dos constantes positivas p y K tales que

$$|\alpha - \alpha_n| \leq K/n^p, \quad \text{para } n \text{ suficientemente grande,}$$

entonces se dice que $\{\alpha_n\}$ **converge a α con orden**, o **velocidad, de convergencia** $\mathcal{O}(1/n^p)$.

Lo anterior se indica escribiendo $\alpha_n = \alpha + \mathcal{O}(1/n^p)$ y diciendo que «$\alpha_n \to \alpha$ con orden de convergencia $1/n^p$». En general lo que interesa es buscar el *mayor* valor de p para el cual $\alpha_n = \alpha + \mathcal{O}(1/n^p)$.

1.4.4. Coste computacional y eficiencia

Normalmente resolvemos un problema en el ordenador mediante un algoritmo, que es una directiva precisa en forma de texto finito, que especifica la ejecución de una serie finita de operaciones elementales. Estamos interesados en aquellos algoritmos que involucran sólo un número finito de etapas.

El *coste computacional* de un algoritmo es el número de operaciones de punto flotante que requiere su ejecución. A menudo la velocidad de un ordenador se mide por el máximo número de operaciones en punto flotante que puede efectuar en un segundo (*flops*, del inglés *floating operation*).

En general, véase [23], el conocimiento exacto del número de operaciones requerido por un algoritmo dado no es esencial. En cambio es útil determinar su orden de magnitud como función de un parámetro δ que está relacionado con la dimensión del problema. Por tanto, decimos que un algoritmo tiene *complejidad constante* si requiere un número de operaciones independiente de δ, es decir $\mathcal{O}(1)$ operaciones, *complejidad lineal* si requiere $\mathcal{O}(\delta)$ operaciones, o, con mayor generalidad, *complejidad polinómica* si requiere $\mathcal{O}(\delta^n)$ operaciones, para un entero positvo n. Otros algoritmos pueden tener *complejidad exponencial*, $\mathcal{O}(k^\delta)$ operaciones (k constante), o incluso *complejidad factorial*, $\mathcal{O}(\delta!)$ operaciones. Aquí el símbolo $\mathcal{O}(\delta^n)$ significa que «se comporta, para δ grande, como una constante multiplicada por δ^n».

Como habitualmente un problema se puede resolver mediante diversos algoritmos, habrá que decidirse por uno. Un criterio útil podría ser el de elegir aquél que proporcione menores errores, pero hay uno mejor (véase [27]), aquél que necesite menor trabajo proporcionando errores dentro de unos límites predeterminados. Esto es lo que se conoce como *eficiencia* de un algoritmo, y depende tanto del tamaño de los errores como del coste computacional del algoritmo.

Terminamos diciendo que otro factor significativo a la hora de analizar los algoritmos suele ser el tiempo necesario para acceder a la memoria del ordenador, que depende de la forma en que el algoritmo esté codificado.

1.5. Sugerencias para seguir leyendo

Para un estudio detallado de la implementación de la aritmética en punto flotante y la adopción del estándar IEEE, véase Higham (2002). Una buena introducción sobre la representación de los números se encuentra en Wilkinson (1994). La mayoría de los textos de cálculo numérico tienen algún tipo de introducción al estudio de errores. Para un punto de vista similar al de este capítulo, véanse: [1, 23], Atkinson (1989) o Stoer y Bulirsch (2002). A modo de complemento, también se puede consultar [22], donde se tratan los fundamentos del cálculo científico de manera extensa.

1.6. Ejercicios

1. Utilícese el polinomio de Taylor de grado 7 de la función $f(x) = \mathrm{e}^x$ alrededor de 0 para aproximar los valores de e^{-3} y $1/\mathrm{e}^3$. ¿Cuál de los dos valores obtenidos da mayor precisión del valor real $\mathrm{e}^{-3} = 4.979 \cdot 10^{-2}$ con 4 cifras significativas? ¿Por qué?

2. Utilícese el polinomio de Taylor de la función $f(x) = \operatorname{sen} x$ alrededor de 0 para aproximar sen(2.3) con una exactitud de 10^{-3}.

3. Determínese la precisión de la aproximación de $\int_1^1 \mathrm{e}^{x^2}\,dx = 1.462651745907$ cuando se reemplaza la función e^{x^2} por su polinomio de Taylor de grado 8 alrededor de 0.

4. Calcúlense los errores absoluto y relativo que se cometen cuando el número $.xyz \times 10^7$ se escribe erróneamente como $x.yz \times 10^7$.

5. ¿Cuántos números de punto flotante hay entre sucesivas potencias de 2?

6. Determínese el número en punto flotante $fl(4/5)$ y calcúlense los errores de redondeo absoluto y relativo que se cometen al representar $4/5$ por $fl(4/5)$.

7. Determínese el número de punto flotante $fl(\sqrt{5})$ utilizando redondeo a 10 cifras significativas.

8. Dados los siguientes valores de x y $\tilde{x}$

   $$a)\ x=\sqrt{5} \text{ y } \tilde{x}=2.2, \quad b)\ x=1/9 \text{ y } \tilde{x}=0.111, \quad c)\ x=\frac{\sqrt{5}}{100} \text{ y } \tilde{x}=0.022, \quad d)\ x=10/9 \text{ y } \tilde{x}=1.11,$$

   determínense los errores absoluto y relativo cuando aproximamos x por $\tilde{x}$.

9. Determínense cuántos números diferentes contiene el conjunto $\mathbb{F}(\beta, t, L, U)$.

10. Determínense cuántos números pertenecen al conjunto $\mathbb{F}(2,2,-2,2)$ y cuál es el valor de ϵ_M para este conjunto.

11. Represéntese en precisión simple, según el estándar IEC559, el número decimal -34.2137.

12. Búsquense ejemplos en los que se ponga de manifiesto que el producto de números en punto flotante no verifica la propiedad asociativa; es decir, $fl(fl(xy)z) \neq fl(xfl(yz))$, para los números de punto flotante x, y, z.

13. La ecuación $x^2 - 100.0001x + 0.01 = 0$ tiene dos soluciones exactas: $x_1 = 100$ y $x_2 = 0.0001$. Obsérvese lo que se obtiene si calculamos x_1 y x_2 con aritmética de punto flotante con 5 decimales mediante las conocidas fórmulas

   $$x_1 = \frac{-b+\sqrt{b^2-4ac}}{2a} \qquad \text{y} \qquad x_2 = \frac{-b-\sqrt{b^2-4ac}}{2a},$$

   que calculan las raíces de la ecuación $ax^2 + bx + c = 0$, $a \neq 0$. ¿Qué ocurre? Calcúlese a continuación la solución x_2 mediante la expresión equivalente

   $$x_2 = \frac{-2c}{b-\sqrt{b^2-4ac}}.$$

   ¿Qué conclusión se puede dar?

14. Justifíquese que si generamos la sucesión $\alpha_n = \left(\frac{1}{9}\right)^n$, $n \in \mathbb{N}$, de forma recursiva mediante $\alpha_0 = 1$ y $\alpha_n = \frac{1}{9}\alpha_{n-1}$, para $n \in \mathbb{N}$, utilizando redondeo a 6 cifras significativas, la sucesión es estable, mientras que si la generamos a partir de $\alpha_0 = 1$, $\alpha_1 = \frac{1}{9}$ y $\alpha_n = \frac{10}{9}\alpha_{n-1} - \alpha_{n-2}$, para $n \geq 2$, la sucesión es inestable.

15. La sucesión de Fibonacci se define mediante

   $$\alpha_0 = 1, \quad \alpha_1 = 1, \quad \alpha_{n+1} = \alpha_n + \alpha_{n-1}, \quad n \in \mathbb{N}.$$

   Demuéstrese que la sucesión $\left\{\dfrac{\alpha_n}{\alpha_{n-1}}\right\}$, $n \in \mathbb{N}$, converge a $\dfrac{1+\sqrt{5}}{2}$.

16. Dadas las sucesiones $\alpha_n = \dfrac{\operatorname{sen} n}{n}$ y $\beta_n = \dfrac{n+3}{n^3}$, $n \in \mathbb{N}$, demuéstrese que la sucesión $\{\beta_n\}$ converge a 0 más rápidamente de lo que lo hace la sucesión $\{\alpha_n\}$.

Capítulo 2

Resolución de sistemas lineales

2.1. Introducción

A menudo tenemos que resolver un sistema de ecuaciones lineales, o simplemente sistema lineal, de la forma

$$A\mathbf{x} = \mathbf{b},$$

donde A es una matriz cuadrada de orden $n \times n$ cuyos elementos a_{ij} son reales o complejos, mientras que $\mathbf{x}$ y $\mathbf{b}$ son dos vectores columna de orden $n \times 1$ con $\mathbf{x}$ representando la solución desconocida y $\mathbf{b}$ un vector dado. Componente a componente, el sistema anterior se puede escribir como

$$\begin{array}{ccccccccc} a_{11}x_1 & + & a_{12}x_2 & + & \cdots & + & a_{1n}x_n & = & b_1, \\ a_{21}x_1 & + & a_{22}x_2 & + & \cdots & + & a_{2n}x_n & = & b_2, \\ & & & & & & & & \vdots \\ a_{n1}x_1 & + & a_{n2}x_2 & + & \cdots & + & a_{nn}x_n & = & b_n. \end{array}$$

La solución del sistema existe si y solo si la matriz A es no singular. En principio, la solución podría calcularse utilizando la conocida *regla de Cramer:*

$$x_i = \frac{\det(A_i)}{\det(A)}, \quad i = 1, 2, \ldots, n,$$

donde A_i es la matriz obtenida a partir de A reemplazando la i-ésima columna por $\mathbf{b}$ y $\det(A)$ denota el determinante de A. Si los $n+1$ determinantes se calculan mediante la conocida *expresión recursiva de Laplace*, se requiere un número total de aproximadamente $2(n+1)!$ operaciones (sumas, restas, productos o divisiones). Por ejemplo, si $n = 24$, se calcularía la suma de 24! sumandos (el número de permutaciones del conjunto $1, 2, \ldots, 24$), y con un ordenador capaz de realizar 10^{15} operaciones en punto flotante por segundo (10^{15} *flops*), se necesitarían 20 años para llevar a cabo este cálculo. El coste computacional es entonces demasiado alto para grandes valores de n que surgen a menudo en las aplicaciones prácticas. Por tanto, hay que recurrir a algoritmos más eficientes.

Pueden utilizarse dos familias de métodos alternativos: los llamados *métodos directos,* que dan la solución del sistema en un número finito de pasos, o los *métodos iterativos,* que requieren (en principio) un número infinito de pasos. La elección entre métodos directos e iterativos puede depender de varios factores: en primer lugar, de la eficiencia teórica del esquema, pero también del tipo particular de matriz, de la memoria de almacenamiento requerida y, finalmente, de la arquitectura del ordenador.

2.2. Métodos directos

Los métodos directos consisten en transformar el sistema $A\mathbf{x} = \mathbf{b}$ en otro equivalente cuya resolución sea prácticamente inmediata. Esta transformación se hace mediante las llamadas operaciones elementales, cuya interpretación matricial nos proporciona una interesante y conocida factorización de la matriz del sistema.

2.2.1. Método de eliminación de Gauss

El método directo de resolución de sistemas de ecuaciones lineales más conocido es el método de eliminación de Gauss, que consiste en transformar el sistema $A\mathbf{x} = \mathbf{b}$, mediante operaciones elementales, en un sistema equivalente triangular superior, cuya resolución es más sencilla y se obtiene mediante la denominada sustitución regresiva o inversa. Recordamos que por operaciones elementales entendemos permutar ecuaciones, multiplicar una ecuación por una constante distinta de cero y sumar a una ecuación una combinación lineal del resto de ecuaciones. Es bien sabido que la solución del sistema no cambia al realizar estas operaciones elementales.

Inicialmente denotamos $A^{(1)} = A$ y $\mathbf{b}^{(1)} = \mathbf{b}$. El método lo podemos interpretar como una sucesión de $n-1$ pasos que dan como resultado una sucesión de matrices y vectores, como se indica a continuación:

$$A^{(1)} \to A^{(2)} \to \cdots \to A^{(n)}, \qquad \mathbf{b}^{(1)} \to \mathbf{b}^{(2)} \to \cdots \to \mathbf{b}^{(n)},$$

de forma que $A^{(n)}$ sea una matriz triangular superior. Así, para el primer paso, supondremos que $a_{11}^{(1)} = a_{11} \neq 0$. Entonces, podemos eliminar x_1 de las $n-1$ últimas ecuaciones. En efecto, tomamos $\ell_{i1} = a_{i1}^{(1)}/a_{11}^{(1)}$ (para $i = 2, \ldots, n$) y sumamos $-\ell_{i1}$ veces la primera ecuación a la i-ésima, esto es

$$a_{ij}^{(2)} = a_{ij}^{(1)} - \ell_{i1}\, a_{1j}^{(1)},\ j = 1, 2, \ldots, n; \quad b_i^{(2)} = b_i^{(1)} - \ell_{i1}\, b_1^{(1)}.$$

(Notemos que $a_{ij}^{(1)}$ designan los elementos de $A^{(1)}$ y $b_i^{(1)}$ son las componentes de $\mathbf{b}^{(1)}$). Si $a_{22}^{(2)} \neq 0$, podemos eliminar de manera similar x_2 de las $n-2$ últimas ecuaciones y así sucesivamente. En general, el paso de la k-ésima matriz a la $(k+1)$-ésima viene dado por las siguientes fórmulas:

$$a_{ij}^{(k+1)} = \begin{cases} a_{ij}^{(k)}, & i = 1, 2, \ldots, k, \\ a_{ij}^{(k)} - \ell_{ik}\, a_{kj}^{(k)}, & i, j = k+1, k+2, \ldots, n, \\ 0, & i = k+1, k+2, \ldots, n;\ j = 1, 2, \ldots, k, \end{cases}$$

donde

$$\ell_{ik} = \begin{cases} \dfrac{a_{ik}^{(k)}}{a_{kk}^{(k)}}, & i = k+1, k+2, \ldots, n, \\ 1, & i = k, \\ 0, & i = 1, 2, \ldots, k-1. \end{cases}$$

Como hemos visto, si los elementos $a_{kk}^{(k)}$ que van apareciendo en los sucesivos pasos (que denominaremos *pivotes*) son no nulos, podemos aplicar el algoritmo con éxito.

Después de a lo sumo n pasos, llegamos al sistema triangular superior $A^{(n)}\mathbf{x} = \mathbf{b}^{(n)}$ siguiente

$$\begin{array}{ccccccccc} a_{11}^{(n)} x_1 & + & a_{12}^{(n)} x_2 & + & \cdots & + & a_{1n}^{(n)} x_n & = & b_1^{(n)}, \\ & & a_{22}^{(n)} x_2 & + & \cdots & + & a_{2n}^{(n)} x_n & = & b_2^{(n)}, \\ & & & & & & \vdots & & \vdots \\ & & & & & & a_{nn}^{(n)} x_n & = & b_n^{(n)}, \end{array}$$

que podemos resolver sin más que llevar a cabo sustitución regresiva

$$\begin{aligned} x_n &= \frac{b_n^{(n)}}{a_{nn}^{(n)}}, \\ x_{n-1} &= \frac{b_{n-1}^{(n)} - a_{n-1,n}^{(n)} x_n}{a_{n-1,n-1}^{(n)}}, \\ &\ \vdots \\ x_k &= \frac{b_k^{(n)} - \sum_{j=k+1}^{n} a_{kj}^{(n)} x_j}{a_{kk}^{(n)}}, \quad \text{para cada} \quad k = n-1, n-2, \ldots, 1. \end{aligned}$$

El procedimiento fallará si en el paso k-ésimo el pivote $a_{kk}^{(k)}$ es cero, porque en ese caso, o bien los *multiplicadores* $\ell_{ik} = a_{ik}^{(k)}/a_{kk}^{(k)}$ no están definidos (lo que ocurre si $a_{kk}^{(k)} = 0$ para algún $k < n$) o no podemos realizar la sustitución regresiva (si $a_{nn}^{(n)} = 0$). Esto no significa que el sistema no tenga solución, sino más bien que la técnica para hallar la solución debe modificarse.

Ejemplo. Vamos a resolver el sistema lineal dado en forma matricial por

$$\begin{pmatrix} 1 & 1 & 1 & 1 \\ 2 & 3 & 1 & 5 \\ -1 & 1 & -5 & 3 \\ 3 & 1 & 7 & -2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 10 \\ 31 \\ -2 \\ 18 \end{pmatrix}.$$

La matriz ampliada junto con los multiplicadores ℓ_{i1} son

$$\begin{matrix} \text{pivote} \rightarrow \\ \ell_{21} = 2 \\ \ell_{31} = -1 \\ \ell_{41} = 3 \end{matrix} \left(\begin{array}{rrrr|r} \boxed{1} & 1 & 1 & 1 & 10 \\ 2 & 3 & 1 & 5 & 31 \\ -1 & 1 & -5 & 3 & -2 \\ 3 & 1 & 7 & -2 & 18 \end{array} \right).$$

Ahora hacemos ceros por debajo del pivote restando múltiplos de la primera ecuación (*primera ecuación pivote*) de las otras tres

$$\begin{matrix} \\ \text{pivote} \rightarrow \\ \ell_{32} = 2 \\ \ell_{42} = -2 \end{matrix} \left(\begin{array}{rrrr|r} 1 & 1 & 1 & 1 & 10 \\ 0 & \boxed{1} & -1 & 3 & 11 \\ 0 & 2 & -4 & 4 & 8 \\ 0 & -2 & 4 & -5 & -12 \end{array} \right).$$

A continuación, hacemos ceros por debajo del nuevo pivote restando múltiplos de la segunda ecuación (*segunda ecuación pivote*) de las dos últimas

$$\begin{matrix} \\ \\ \text{pivote} \rightarrow \\ \ell_{43} = -1 \end{matrix} \left(\begin{array}{rrrr|r} 1 & 1 & 1 & 1 & 10 \\ 0 & 1 & -1 & 3 & 11 \\ 0 & 0 & \boxed{-2} & -2 & -14 \\ 0 & 0 & 2 & 1 & 10 \end{array} \right).$$

Restamos finalmente múltiplos de la tercera ecuación (*tercera ecuación pivote*) de la última para hacer ceros por debajo del último pivote y obtener así el siguiente sistema triangular superior:

$$\left(\begin{array}{rrrr|r} 1 & 1 & 1 & 1 & 10 \\ 0 & 1 & -1 & 3 & 11 \\ 0 & 0 & -2 & -2 & -14 \\ 0 & 0 & 0 & -1 & -4 \end{array} \right).$$

Para terminar, aplicamos sustitución regresiva al sistema anterior y obtenemos:

$$x_4 = 4, \quad x_3 = \frac{-14 + 2x_4}{-2} = 3, \quad x_2 = 11 + x_3 - 3x_4 = 2, \quad x_1 = 10 - x_2 - x_3 - x_4 = 1. \quad \square$$

Nota. Obsérvese en el ejemplo anterior que podemos utilizar los pasos que se dan para resolver el sistema lineal $A\mathbf{x} = \mathbf{b}$ por el método de eliminación de Gauss para hallar la matriz triangular superior

$$U = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & -1 & 3 \\ 0 & 0 & -2 & -2 \\ 0 & 0 & 0 & -1 \end{pmatrix}$$

de la matriz

$$A = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 2 & 3 & 1 & 5 \\ -1 & 1 & -5 & 3 \\ 3 & 1 & 7 & -2 \end{pmatrix}.$$

2.2.2. Estrategias de pivoteo

En el algoritmo de eliminación de Gauss hemos supuesto que todos los pivotes son no nulos. Pero esto no ocurre en muchas ocasiones, o aunque ocurra, si el pivote es muy pequeño, puede provocar resultados inexactos, como puede comprobarse en los siguientes dos ejemplos que aparecen en [17]:

$$\left\{\begin{array}{rcrcl} & & y & = & 1, \\ x & + & y & = & 2, \end{array}\right. \qquad \left\{\begin{array}{rcrcl} \varepsilon x & + & y & = & 1, \\ x & + & y & = & 2, \end{array}\right.$$

donde ε es un número muy pequeño pero distinto de cero. Las estrategias de pivoteo evitan parcialmente estos problemas.

A partir de los dos ejemplos anteriores y otros parecidos se puede extraer la idea de que el intercambio de ecuaciones (filas) en un sistema, cuando la situación lo requiera, es una buena estrategia. Así, si $a_{kk}^{(k)} = 0$, entonces, podemos buscar $a_{ik}^{(k)} \neq 0$ con $i > k$ (lo cual es siempre posible ya que $\det(\mathrm{A}) \neq 0$), intercambiar las filas k-ésima e i-ésima y continuar el proceso. Este intercambio de filas se conoce como *pivoteo.*

Si $a_{kk}^{(k)} \neq 0$ pero es pequeño, aunque el método de Gauss se puede aplicar, es muy inestable numéricamente, con lo que pequeños errores en los datos pueden originar grandes cambios en la solución (véase el ejemplo 1 de la pág. 268 de [7]). En este caso, también se tendrá que pivotar.

En el método general de eliminación de Gauss se puede utilizar una ecuación (o fila) como ecuación pivote (o fila pivote) solo si el elemento pivote no es cero. Si el elemento pivote es cero, la ecuación (fila) se puede intercambiar con una de las ecuaciones (filas) que están por debajo que tenga un elemento pivote no cero. Las opciones de pivoteo más habituales que se utilizan en el paso k-ésimo son:

- *Pivoteo parcial:* se toma como pivote el elemento de mayor módulo de entre los $n-k$ últimos elementos de la columna k-ésima; es decir, se elige $a_{ik}^{(k)}$ $(i = k, k+1, \ldots, n)$ de forma que

  $$\left|a_{ik}^{(k)}\right| = \max_{k\leq j\leq n} \left|a_{jk}^{(k)}\right|,$$

  e intercambiamos las filas i-ésima y k-ésima.

- *Pivoteo total:* se toma como pivote el elemento de mayor módulo de la submatriz correspondiente de la matriz $A^{(k)}$; es decir, se elige el elemento $a_{ij}^{(k)}$ $(i, j = k, k+1, \ldots, n)$ de modo que

  $$\left|a_{ij}^{(k)}\right| = \max_{k\leq r,s\leq n} \left|a_{rs}^{(k)}\right|,$$

  e intercambiamos las filas y columnas correspondientes. En este pivoteo, si el pivote elegido no está en la columna k-ésima, el intercambio de columnas que hay que realizar implica intercambiar el orden de las incógnitas, lo que habrá que tener en cuenta a la hora de resolver el sistema lineal.

Como con el pivoteo parcial solo hay que intercambiar filas de la matriz, mientras que con el pivoteo total también hay que intercambiar columnas, el coste operacional de este último es mucho más elevado. En general, esto hace que en el método de eliminación de Gauss se utilice habitualmente la estrategia de pivoteo parcial.

COMENTARIO ADICIONAL. ▷ En el ejemplo anterior hemos podido aplicar el método de eliminación de Gauss sin necesidad de pivoteo, pero ¿cuándo podemos asegurar que no será necesario hacerlo? Digamos, en este caso, que si la matriz A es **simétrica definida positiva** (es decir, $A = A^T$ y tal que $\mathbf{x}^T A\mathbf{x} > 0$, para todo vector columna $\mathbf{x} \neq 0$) o **estrictamente diagonal dominante** (esto es que los elementos de A cumplen que $|a_{ii}| > \sum_{j=1,j\neq i}^{n} |a_{ij}|$ para $i = 1, 2, \ldots, n$), el algoritmo de Gauss es numéricamente estable y se puede llevar a cabo sin necesidad de pivoteo. Dos equivalencias que hacen más fácil, en algunos casos, ver que una matriz simétrica es definida positiva son: una, que todos sus valores propios son > 0; y dos, que sus n menores principales son > 0.

2.2.3. Métodos de factorización

Hemos visto que el método de eliminación de Gauss consiste de dos partes. La primera parte es el procedimiento de eliminación en el que el sistema de ecuaciones lineales $A\mathbf{x} = \mathbf{b}$ se transforma en un sistema

equivalente de ecuaciones $A'\mathbf{x} = \mathbf{b}'$, donde la nueva matriz de coeficientes A' es triangular superior. En la segunda parte, el sistema equivalente se resuelve mediante sustitución regresiva. El procedimiento de eliminación requiere de muchas operaciones matemáticas y significativamente más tiempo de computación que los cálculos de la sustitución regresiva. Durante el procedimiento de elimanción la matriz de coeficientes A y el vector $\mathbf{b}$ cambian, lo que significa que si tenemos la necesidad de resolver varios sistemas de ecuaciones que tengan la misma matriz de coeficientes A, pero diferentes vectores $\mathbf{b}$, el procedimiento de eliminación tiene que realizarse para cada uno de los vectores $\mathbf{b}$. Esto se puede mejorar si las operaciones asociadas a la matriz A se disocian de las operaciones asociadas al vector $\mathbf{b}$. De esta forma, el procedimiento de elimación con A se hace una sola vez y se utiliza entonces para resolver los sistemas de ecuaciones con distintos $\mathbf{b}$.

Una opción para resolver varios sistemas de ecuaciones de la forma $A\mathbf{x} = \mathbf{b}$ que tengan la misma matriz A pero diferentes vectores $\mathbf{b}$ es calcular primero la matriz inversa A^{-1} de A y, una vez que esta matriz inversa es conocida, la solución se puede calcular así: $\mathbf{x} = A^{-1}\mathbf{b}$. Sin embargo, el cálculo de la matriz inversa requiere de muchas operaciones matemáticas y computacionalmente es ineficiente. Un método más eficiente para calcular la solución $\mathbf{x}$ es el **método de factorización LU.**

En el método de factorización LU las operaciones con la matriz A se realizan sin utilizar, o cambiar, el verctor $\mathbf{b}$, que se utiliza solo en la parte de sustitución de la solución. En este caso, sea A una matriz cuadrada de orden n y supongamos que existen dos matrices convenientes L y U, triangular inferior y triangular superior, respectivamente, tales que $A = LU$. Llamamos a la igualdad anterior una *factorización LU* (o descomposición) de A. Si A es no singular, lo mismo ocurre con L y U, y de este modo sus elementos diagonales son no nulos.

En tal caso, resolver $A\mathbf{x} = \mathbf{b}$ conduce a la solución de los dos sistemas triangulares $L\mathbf{z} = \mathbf{b}$ y $U\mathbf{x} = \mathbf{z}$. Se puede proceder por etapas como sigue:

1. Resolución de $L\mathbf{z} = \mathbf{b}$ para $\mathbf{z}$ por sustitución progresiva:

$$\begin{aligned} z_1 &= \frac{b_1}{\ell_{11}}, \\ z_k &= \frac{1}{\ell_{kk}}(b_k - \ell_{k1}z_1 - \ell_{k2}z_2 - \cdots - \ell_{k,k-1}z_{k-1}), \quad k = 2, 3, \ldots, n. \end{aligned}$$

2. Resolución de $U\mathbf{x} = \mathbf{z}$ para $\mathbf{x}$ por sustitución regresiva:

$$\begin{aligned} x_n &= \frac{z_n}{u_{nn}}, \\ x_k &= \frac{1}{u_{kk}}(z_k - u_{k,k+1}x_{k+1} - u_{k,k+2}x_{k+2} \cdots - u_{kn}x_n), \quad k = k-1, k-2, \ldots, 1. \end{aligned}$$

En este punto necesitamos un algoritmo que permita un cálculo efectivo de los factores L y U de la matriz A.

La factorización anterior se puede obtener simplemente mediante la fórmula del producto de matrices, que conduce a un sistema lineal de n^2 ecuaciones en el que las incógnitas son los $n^2 + n$ coeficientes de las matrices triangulares L y U. Así, si $A = (a_{ij})$, $L = (\ell_{ij})$ y $U = (u_{ij})$, donde los elementos ℓ_{ij} y u_{ij} están por determinar, realizando el producto se tiene que

$$a_{ij} = \sum_{p=1}^{\text{mín}(i,j)} \ell_{ip}\, u_{pj}, \quad i, j = 1, 2, \ldots, n,$$

donde la última igualdad se debe a que $\ell_{ip} = 0$, para $p > i$, y $u_{pj} = 0$, para $p > j$. Entonces, si fijamos de antemano el valor de ℓ_{ii} o de u_{ii} (distinto de cero), para todo i, la igualdad anterior permite el cálculo de L y U, dando lugar a formas compactas de factorización.

Factorización de Doolittle

Si se fija $\ell_{ii} = 1$, para todo i, la factorización correspondiente se denomina *factorización de Doolittle,* o simplemente, *factorización LU.*

Si podemos aplicar el método de eliminación de Gauss sin pivoteo a la matriz A, teniendo cuidado de guardar los mulplicadores ℓ_{ik} en una matriz L con elementos 1 en su diagonal, y llamamos U a la matriz triangular superior obtenida tras la aplicación del método de Gauss, se tiene que $A = LU$.

Ejemplo. En la matriz A del ejemplo anterior se tiene que la matriz L de la factorización $LU = A$ se construye a partir de los coeficientes ℓ_{ik}, obtenidos en dicho ejemplo, de la siguiente forma

$$L = \begin{pmatrix} 1 & 0 & 0 & 0 \\ \ell_{21} & 1 & 0 & 0 \\ \ell_{31} & \ell_{32} & 1 & 0 \\ \ell_{41} & \ell_{42} & \ell_{43} & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ -1 & 2 & 1 & 0 \\ 3 & -2 & -1 & 1 \end{pmatrix}.$$

Comprobamos que la factorización es correcta

$$LU = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ -1 & 2 & 1 & 0 \\ 3 & -2 & -1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & -1 & 3 \\ 0 & 0 & -2 & -2 \\ 0 & 0 & 0 & -1 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 2 & 3 & 1 & 5 \\ -1 & 1 & -5 & 3 \\ 3 & 1 & 7 & -2 \end{pmatrix} = A. \quad \square$$

Comentario adicional. ▷ Una matriz A tiene factorización de Doolittle si y solo si se le puede aplicar el método de eliminación de Gauss sin pivoteo.

Factorización de Crout

La *factorización de Crout* es similar a la Doolittle, pero con la diferencia de que ahora fijamos $u_{ii} = 1$, para todo i, de manera que la matriz U es la que tiene los elementos diagonales iguales a 1. Véase [12] para mayor detalle.

Factorización LU con pivoteo (o factorización $PA = LU$)

Para realizar las factorizaciones anteriores de Doolittle y Crout, hemos supuesto que es posible realizar todos los cálculos sin pivoteo. Si se utiliza pivoteo, entonces las matrices L y U que se obtienen no son la factorización de la matriz original A. El producto LU da una matriz con filas que tienen los mismos elementos que A, pero como consecuencia del pivoteo, las filas están en un orden diferente. Cuando se utiliza el intercambio de filas como operación elemental en el método de Gauss, éste lleva consigo una modificación menor en la factorización LU, y los cambios que se han hecho tienen que ser grabados y almacenados. Esto se puede hacer creando una matriz P, llamada *matriz de permutación*, tal que $PA = LU$ y donde P es el resultado de permutar filas en la matriz identidad. En este caso, P contiene información sobre el orden que impone el método de eliminación de Gauss con pivoteo, y por tanto es equivalente a reordenar la matriz A de acuerdo a ese orden y aplicar luego el método de eliminación de Gauss sin pivoteo. Si las matrices L y U se utilizan para resolver un sistema de ecuaciones $A\mathbf{x} = \mathbf{b}$, entonces hay que cambiar el orden de las filas de $\mathbf{b}$ para que sea consistente con el pivoteo. Esto se consigue multiplicando $\mathbf{b}$ por la matriz de permutaicón, $P\mathbf{b}$. Para mayor detalle puede consultarse [14].

Alternativamente, véase [6], la factorización $PA = LU$ se puede escribir como $A = \widehat{L}U$, donde $\widehat{L} = P^{-1}L$, de forma que $\widehat{L}$ es una permutación de la matriz triangular inferior (que es el resultado de reordenar las filas de L).

Ejemplo. La matriz

$$A = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 2 \\ 1 & 2 & 2 \end{pmatrix}$$

tiene el menor principal 2×2 nulo, luego no tiene factorización de Doolitte y, por tanto, no es posible aplicar el método de eliminación de Gauss sin pivoteo. Es necesario permutar las filas 2 y 3 para llevar a cabo el método de eliminación de Gauss; es decir, se necesita pivoteo. Si multiplicamos entonces la matriz A por la matriz de permutación de filas 2 y 3

$$P = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \quad \text{se tiene la matriz} \quad PA = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 1 & 1 & 2 \end{pmatrix},$$

que ya no necesita pivoteo, puesto que los tres menores principales de PA son mayores que cero. $\square$

Factorización de Cholesky

El algoritmo para la *factorización de Cholesky* es el caso especial del algoritmo general para la factorización LU en el que $U = L^T$, de modo que $\ell_{ii} = u_{ii}$, para $i = 1, 2, \ldots, n$. Si A es simétrica definida positiva, entonces tiene una factorización única de la forma $A = LL^T$, donde L es triangular inferior con todos los elementos de la diagonal positivos. Como

$$a_{ij} = \sum_{p=1}^{n} \ell_{ip}\, u_{pj} = \sum_{p=1}^{\min(i,j)} \ell_{ip}\, \ell_{jp} \qquad \Longrightarrow \qquad \ell_{kk} = \left(a_{kk} - \sum_{p=1}^{k-1} \ell_{kp}^2 \right)^{\frac{1}{2}},$$

y los valores ℓ_{kk} quedan determinados. Así, los elementos de la matriz L se obtienen de la siguiente forma. Para $k = 1, 2, \ldots, n$, se tiene:

$$\ell_{kk} = \left(a_{kk} - \sum_{p=1}^{k-1} \ell_{kp}^2 \right)^{\frac{1}{2}}, \quad \ell_{ik} = \frac{1}{\ell_{kk}} \left(a_{ik} - \sum_{p=1}^{k-1} \ell_{ip}\, \ell_{kp} \right), \quad i = k+1, k+2, \ldots, n.$$

Estas igualdades pueden deducirse escribiendo la igualdad $A = LL^T$ como un sistema de ecuaciones.

Ejemplo. Factorizaremos mediante el método de Cholesky la siguiente matriz

$$A = \begin{pmatrix} 16 & 4 & 4 \\ 4 & 26 & 6 \\ 4 & 6 & 11 \end{pmatrix}.$$

De las fórmulas anteriores, obtenemos

$$\ell_{11} = 4; \qquad \ell_{21} = 1, \quad \ell_{22} = 5; \qquad \ell_{31} = 1, \quad \ell_{32} = 1, \quad \ell_{33} = 3.$$

Luego,

$$L = \begin{pmatrix} 4 & 0 & 0 \\ 1 & 5 & 0 \\ 1 & 1 & 3 \end{pmatrix} \qquad \text{y} \qquad U = L^T = \begin{pmatrix} 4 & 1 & 1 \\ 0 & 5 & 1 \\ 0 & 0 & 3 \end{pmatrix}. \quad \square$$

Comentario adicional. ▷ Los métodos de factorización LU son particularmente útiles cuando se necesita resolver un conjunto de sistemas $A\mathbf{x_j} = \mathbf{b_j}$, $j = 1, 2, \ldots, n$, que tienen la misma matriz cuadrada de coeficientes A. En muchos problemas de la ciencia y la ingeniería aparecen sistemas de este tipo y, en este caso, es más eficiente utilizar métodos LU que aplicar separadamente el método de eliminación de Gauss a cada uno de los n sistemas. Este método es especialmente útil para calcular la matriz inversa de A (véase ejercicio el 5). Además, se puede economizar el espacio de almacenamiento al no necesitar almacenar los ceros de L o U y los unos de la diagonal de L o U (según sea la factorización), de manera que las matrices L y U se construyen almacenando sus elementos en el espacio de A.

2.3. Métodos iterativos

Los métodos estudiados hasta ahora se llaman métodos directos porque encuentran la solución exacta tras un número finito de pasos (salvo errores de redondeo). A continuación estudiaremos otros métodos, llamados iterativos, que si bien necesitarían un número infinito de pasos para alcanzar la solución, permiten aproximarla tras un número finito. Además están mejor adaptados a la resolución de grandes sistemas lineales, sobre todo si estos están dados por matrices con un alto porcentaje de elementos nulos (*matrices dispersas*).

La idea básica de un método iterativo para resolver el sistema $A\mathbf{x} = \mathbf{b}$ consiste en construir una sucesión de vectores $\left\{ \mathbf{x}^{(k)} = \left(x_1^{(k)}, x_2^{(k)}, \ldots, x_n^{(k)} \right)^T;\ k \in \mathbb{N} \right\}$ de $\mathbb{R}^n$ que *converja* a la solución exacta $\mathbf{x}$, esto es

$$\lim_{k \to \infty} \mathbf{x}^{(k)} = \mathbf{x},$$

para un vector inicial dado $\mathbf{x}^{(0)}$ de $\mathbb{R}^n$.

Una técnica general para construir un método iterativo se basa en una *descomposición* de la matriz A del sistema $A\mathbf{x} = \mathbf{b}$ en la forma $A = M - (M - A)$, donde M es una matriz no singular adecuada (llamada el *precondicionador* de A). Entonces

$$M\mathbf{x} = (M - A)\mathbf{x} + \mathbf{b},$$

que tiene la forma $\mathbf{x} = T\mathbf{x} + \mathbf{c}$, donde $T = M^{-1}(M - A)$ es una matriz $n \times n$, llamada *matriz de iteración*, y $\mathbf{c} = M^{-1}\mathbf{b}$. En correspondencia con esta descomposición y después de seleccionar un vector inicial $\mathbf{x}^{(0)}$, podemos definir el siguiente método iterativo:

$$\mathbf{x}^{(k)} = T\mathbf{x}^{(k-1)} + \mathbf{c}, \quad k \in \mathbb{N}.$$

Si el proceso converge, el límite es solución de la ecuación planteada y, por tanto, solución del sistema inicial. El sistema inicial es fácil de resolver si M es fácil de invertir, en el sentido de que sea fácil resolver un sistema asociado a dicha matriz, como por ejemplo ocurre cuando M es una matriz diagonal o triangular.

A continuación nos planteamos cuándo la sucesión anterior convergerá a la solución. Si denotamos el error cometido en la k-ésima iteración por $\mathbf{e}^{(k)} = \mathbf{x}^{(k)} - \mathbf{x}$, se verifica que

$$\mathbf{e}^{(k)} = \mathbf{x}^{(k)} - \mathbf{x} = (T\mathbf{x}^{(k-1)} + \mathbf{c}) - (T\mathbf{x} + \mathbf{c}) = T(\mathbf{x}^{(k-1)} - \mathbf{x}) = T\mathbf{e}^{(k-1)}$$

y, por tanto,

$$\mathbf{e}^{(k)} = T\mathbf{e}^{(k-1)} = T^2\mathbf{e}^{(k-2)} = \cdots = T^k\mathbf{e}^{(0)}, \quad k \in \mathbb{N}.$$

Así, el error en las iteraciones depende fundamentalmente de las potencias sucesivas de la matriz T. A continuación, eligiendo de forma conveniente cualesquier norma vectorial y norma matricial subordinada, llegamos a la desigualdad

$$\|\mathbf{e}^{(k)}\| \leq \|T\|^k \|\mathbf{e}^{(0)}\|.$$

Entonces, si $\|T\| < 1$, podemos concluir de inmediato que $\mathbf{e}^{(k)} \to 0$ cuando $k \to \infty$ para cada posible $\mathbf{e}^{(0)}$ (y por tanto $\mathbf{x}^{(0)}$). A continuación damos un criterio en el que se pone de manifiesto que esta propiedad es necesaria para la convergencia.

El **criterio fundamental de convergencia** de los métodos iterativos solo involucra la matriz T del método iterativo considerado y establece que son equivalentes ([14]):

a) el método asociado a la matriz T es convergente,

b) $\rho(T) < 1$, donde $\rho(T)$ es el *radio espectral* de la matriz T, esto es,

$$\rho(T) = \text{máx}\{|\lambda|;\ \lambda \text{ es valor propio de } T\},$$

c) $\|T\| < 1$ para alguna norma matricial.

Las normas matriciales subordinadas son normas que provienen de una norma vectorial. Las normas $||\cdot||_1$ y $||\cdot||_\infty$ son ejemplos de tales normas (véanse sus definiciones en el capítulo anterior).

Comentarios adicionales. ▷ A la hora de resolver un sistema lineal mediante un método iterativo, en primer lugar deberemos asegurar su convergencia (por ejemplo, encontrando alguna norma para la cual $\|T\| < 1$ o viendo que $\rho(T) < 1$). A continuación, y en caso de disponer de varios métodos a nuestro alcance, elegiremos aquél cuya matriz asociada tenga una norma o un radio espectral menor.

▷ Como en un método iterativo no podemos esperar hallar la solución exacta, sino calcular una aproximación, debemos fijar un **criterio de parada** que dé por terminado el método cuando la aproximación obtenida se considere suficientemente buena. Una posibilidad es medir la diferencia entre dos iteraciones consecutivas e iterar hasta que

$$\left\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\right\|$$

sea tan pequeño como una cierta cantidad prescrita $\epsilon > 0$, que se denomina *tolerancia*, de manera que se considera que se está suficientemente cerca de la solución y se finaliza el método. Otros criterios se basan en la diferencia relativa entre dos iteraciones consecutivas (véase la pág. 157 de [16]). Ahora bien, según el criterio fundamental de convergencia, la convergencia o divergencia del método iterativo solo depende del carácter de las propias matrices, pero en el caso de que haya convergencia, una buena aproximación inicial hará que el número de iteraciones sea relativamente pequeño.

▷ Desde un punto de vista computacional, es importante evitar que el método entre en un bucle infinito. Para ello, se suele fijar un número máximo de iteraciones de forma que si se supera, se termine el método con un mensaje de error. (Existe también otro problema: que la solución crezca de forma que supere la cantidad máxima representable en punto flotante (*overflow*)).

2.3.1. Los métodos de Jacobi y Gauss-Seidel

Si los elementos diagonales de A son todos distintos de cero, podemos escribir

$$A = D - L - U,$$

donde D es la matriz diagonal que contiene los elementos diagonales de A,

$$L = \begin{pmatrix} 0 & \dots & 0 & 0 \\ -a_{21} & \ddots & \vdots & \vdots \\ \vdots & \ddots & 0 & \vdots \\ -a_{n1} & \cdots & -a_{n,n-1} & 0 \end{pmatrix} \quad \text{y} \quad U = \begin{pmatrix} 0 & -a_{12} & \dots & -a_{1n} \\ \vdots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & -a_{n-1,n} \\ 0 & \cdots & \cdots & 0 \end{pmatrix}.$$

La ecuación $A\mathbf{x} = \mathbf{b}$ o $(D - L - U)\mathbf{x} = \mathbf{b}$ se transforma entonces en

$$D\mathbf{x} = (L + U)\mathbf{x} + \mathbf{b},$$

y, si existe D^{-1} (es decir, $a_{ii} \neq 0$ para todo i), entonces

$$\mathbf{x} = D^{-1}(L + U)\mathbf{x} + D^{-1}\mathbf{b}.$$

De donde se obtiene la expresión matricial del **método iterativo de Jacobi:**

$$\mathbf{x}^{(k)} = T_j\, \mathbf{x}^{(k-1)} + \mathbf{c}_j, \quad \text{para cada} \quad k \in \mathbb{N},$$

con $T_j = D^{-1}(L + U)$ y $\mathbf{c}_j = D^{-1}\mathbf{b}$.

EJEMPLO. Resolvemos el sistema

$$\begin{array}{rcrcrcrcr} 7x_1 & - & 2x_2 & + & x_3 & & & = & 17 \\ x_1 & - & 9x_2 & + & 3x_3 & - & x_4 & = & 13 \\ 2x_1 & & & + & 10x_3 & + & x_4 & = & 15 \\ x_1 & - & x_2 & + & x_3 & + & 6x_4 & = & 10 \end{array}$$

mediante el método de Jacobi, utilizando una tolerancia $\epsilon = 10^{-3}$, un número máximo de 30 iteraciones y el vector inicial $\mathbf{x}^{(0)} = \mathbf{0} = (0, 0, 0, 0)^T$.

En primer lugar, reescribimos las ecuaciones anteriores de la forma

$$\begin{aligned} x_1 &= \tfrac{1}{7}(17 + 2x_2 - x_3) \\ x_2 &= \tfrac{1}{9}(-13 + x_1 + 3x_3 - x_4) \\ x_3 &= \tfrac{1}{10}(15 - 2x_1 - x_4) \\ x_4 &= \tfrac{1}{6}(10 - x_1 + x_2 - x_3), \end{aligned}$$

que dan el siguiente método iterativo de Jacobi

$$\begin{aligned} x_1^{(k+1)} &= \tfrac{1}{7}\left(17 + 2x_2^{(k)} - x_3^{(k)}\right) \\ x_2^{(k+1)} &= \tfrac{1}{9}\left(-13 + x_1^{(k)} + 3x_3^{(k)} - x_4^{(k)}\right) \\ x_3^{(k+1)} &= \tfrac{1}{10}\left(15 - 2x_1^{(k)} - x_4^{(k)}\right) \\ x_4^{(k+1)} &= \tfrac{1}{6}\left(10 - x_1^{(k)} + x_2^{(k)} - x_3^{(k)}\right). \end{aligned}$$

Sustituyendo $\mathbf{x}^{(0)} = (0,0,0,0)^T$ en el lado derecho de cada una de las ecuaciones obtenemos

$$\begin{aligned} x_1^{(1)} &= \tfrac{1}{7}(17 + 2(0) - 0) = 2.428571429 \\ x_2^{(1)} &= \tfrac{1}{9}(-13 + 0 + 3(0) - 0) = -1.444444444 \\ x_3^{(1)} &= \tfrac{1}{10}(15 - 2(0) - 0) = 1.5 \\ x_4^{(1)} &= \tfrac{1}{6}(10 - 0 + 0 - 0) = 1.666666667. \end{aligned}$$

Luego, $\mathbf{x}^{(1)} = (2.428571429, -1.444444444, 1.5, 1.666666667)^T$. Procediendo de la misma manera, véase [16], se genera una sucesión convergente a

$$\mathbf{x}^{(9)} = (2.000127203, -1.000100162, 1.000118096, 1.000162172)^T$$

con una tolerancia de 10^{-3} y utilizando la norma del máximo $\|\mathbf{x}\|_\infty = \max\limits_{i=1,2,3,4} |x_i|$. $\square$

A continuación, razónese que si escribimos $\mathbf{x}^{(k)} = \left(x_1^{(k)}, x_2^{(k)}, \ldots, x_n^{(k)}\right)^T$, se tiene que el valor de cada componente de la iteración, para $k \in \mathbb{N}$, es:

$$x_i^{(k)} = \frac{1}{a_{ii}} \left(- \sum_{j=1,\ j\neq i}^{n} \left(a_{ij}\, x_j^{(k-1)} \right) + b_i \right), \quad \text{para cada} \quad i = 1, 2, \ldots, n.$$

Reconsiderando la última igualdad podemos ver algo que seguramente mejora el método de Jacobi. En dicha ecuación se utilizan todas las componentes de $\mathbf{x}^{(k-1)}$ para calcular $x_i^{(k)}$. Puesto que las componentes $x_1^{(k)}, x_2^{(k)}, \ldots, x_{i-1}^{(k)}$ de $\mathbf{x}^{(k)}$ ya se han calculado y son probablemente mejores aproximaciones de las soluciones exactas $x_1, x_2, \ldots, x_{i-1}$ que $x_1^{(k-1)}, x_2^{(k-1)}, \ldots, x_{i-1}^{(k-1)}$, podemos calcular $x_i^{(k)}$ usando estos valores calculados más recientemente; es decir, podemos utilizar

$$x_i^{(k)} = \frac{1}{a_{ii}} \left(- \sum_{j=1}^{i-1} \left(a_{ij} x_j^{(k)} \right) - \sum_{j=i+1}^{n} \left(a_{ij} x_j^{(k-1)} \right) + b_i \right), \quad \text{para cada} \quad i = 1, 2, \ldots, n.$$

Esta modificación recibe el nombre de **método iterativo de Gauss-Seidel**.

Con las definiciones de D, L y U que hemos usado antes, razónese que el método de Gauss-Seidel se puede representar mediante

$$\mathbf{x}^{(k)} = T_g\, \mathbf{x}^{(k-1)} + \mathbf{c}_g, \quad \text{para cada} \quad k \in \mathbb{N},$$

donde $T_g = (D - L)^{-1} U$ y $\mathbf{c}_g = (D - L)^{-1}\mathbf{b}$. Como $\det(D - L) = a_{11} a_{22} \cdots a_{22}$, la matriz triangular inferior $D - L$ es invertible precisamente si $a_{ii} \neq 0$ para cada $i = 1, 2, \ldots, n$.

Comentario adicional. ▷ Como consecuencia de que los nuevos valores se pueden almacenar inmediatamente en el lugar de los valores antiguos, los requerimientos de almacenamiento para $\mathbf{x}$ con el método de Gauss-Seidel son la mitad de lo que serían para el método de Jacobi.

2.3.2. Acerca de la convergencia de los métodos de Jacobi y Gauss-Seidel

Por un lado, según el criterio fundamental de convergencia, los métodos iterativos de Jacobi y de Gauss-Seidel convergen si y solo si $\rho(T_j) < 1$ y $\rho(T_g) < 1$. Esto es una condición necesaria y suficiente, pero es necesario calcular el radio espectral, lo que puede ser tan costoso de calcular como resolver el sistema. Sin embargo, existen condiciones suficientes que aseguran de una forma más sencilla la convergencia de ambos métodos.

Resultados de convergencia

Por un lado, podemos establecer las siguientes propiedades de convergencia *a priori* para los métodos de Jacobi y de Gauss-Seidel:

- Si la matriz A es estrictamente diagonal dominante, entonces ambos métodos convergen para cualquier elección inicial $\mathbf{x}^{(0)}$.
- Si las matrices A y $2D - A$ son simétricas definidas positivas, entonces el método de Jacobi converge para todo $\mathbf{x}^{(0)}$.
- Si la matriz A es simétrica definida positiva, entonces el método de Gauss-Seidel converge para todo $\mathbf{x}^{(0)}$.

Por otro lado, la sección anterior parece indicar que el método de Gauss-Seidel es superior al método de Jacobi, pero no existen resultados generales que muestren que el método de Gauss-Seidel converge más rápido que el de Jacobi. Esto es verdad casi siempre, pero hay sistemas lineales para los que el método de Jacobi converge, mientras que el de Gauss-Seidel no. En algunos casos particulares si que se puede dar la superioridad del método de Gauss-Seidel sobre el método de Jacobi, como se establece a continuación:

Si la matriz A es *tridiagonal* (matriz con elmentos no nulos en a lo sumo la diagonal principal y las diagonales inmediatamente superior e inmediantemente inferior), no singular y con todos elementos diagonales distintos de cero, entonces los métodos de Jacobi y Gauss-Seidel son ambos divergentes o son ambos convergentes. En el último caso, los radios espectrales de las matrices T_j y T_g verifican que $\rho(T_j)^2 = \rho(T_g)$.

Mejora de la convergencia mediante relajación

Podemos modificar el método de Gauss-Seidel introduciendo un parámetro ω que permita acelerar la convergencia del método. Después de calcular cada nuevo valor de $\mathbf{x}$, mediante el método de Gauss-Seidel, modificamos ese valor mediante una combinación lineal de los resultados de las iteraciones anterior y actual:

$$x_i^{(\text{nuevo})} = \omega\, x_i^{(\text{nuevo})} + (1-\omega)\, x_i^{(\text{anterior})}, \quad \text{para cada} \quad i = 1, 2, \ldots, n.$$

donde ω puede tomar valores comprendidos entre 0 y 2. Para $0 < \omega < 1$, se llama método de subrelajación sucesiva y, para $1 < \omega$, se llama método de sobrerelajación sucesiva (o método SOR, del inglés *successive over relaxation*). Obsérvese que el método SOR se reduce al método de Gauss-Seidel si $\omega = 1$.

Damos a continuación algunas propiedades de convergencia *a priori* para el método SOR:

- Si la matriz A es simétrica definida positiva y $0 < \omega < 2$, entonces el método SOR converge para todo $\mathbf{x}^{(0)}$.
- Si la matriz A es estrictamente diagonal dominante y $0 < \omega \leq 1$, entonces el método SOR converge para todo $\mathbf{x}^{(0)}$.

Los resultados anteriores tratan la convergencia del método SOR, pero no dicen nada acerca de la elección del parámetro ω. Con relación a esto último, podemos decir que se puede elegir de forma óptima el parámetro ω para un caso particular de matrices:

Si la matriz A es simétrica definida positiva y tridiagonal, entonces el valor óptimo del parámetro ω es

$$\omega = \frac{2}{1 + \sqrt{1 - \rho(T_j)^2}}.$$

Notemos que las matrices tridiagonales aparecen frecuentemente cuando se trabaja con métodos de interpolación mediante *splines* cúbicos y métodos de diferencias finitas para resolver problemas de contorno y ecuaciones en derivadas parciales.

2.4. Sugerencias para seguir leyendo

Los tópicos introducidos en este capítulo son parte del campo del álgebra lineal numérica, que también incluye los tópicos tratados en el complemento A. Un estudio interesante sobre la estrategia de pivoteo se encuentra en Hager (1998). Una discusión sobre el escalado en el pivoteo y el cálculo sobre el número de operaciones de los diferentes métodos puede consultarse en Atkinson (1989). Dos referencias a considerar para la resolución de sistemas lineales son Wilkinson (1994) para los métodos directos y Varga (2000) para los métodos iterativos. Para más información sobre la computación matricial, véase Golub y Van Loan (1996).

2.5. Ejercicios

1. Resuélvanse los siguientes sistemas por el método de eliminación de Gauss

$$a)\left\{\begin{array}{l} x+y=4\\ 2x+2z=4\\ 3y+3z=4\end{array}\right. \qquad b)\left\{\begin{array}{l} 0.15x+2.11y+30.75z=-26.38\\ 0.64x+1.21y+2.05z=1.01\\ 3.21x+1.53y+1.04z=5.23\end{array}\right.$$

   sin pivoteo, con pivoteo parcial y con pivoteo total.

2. Resuélvanse los sistemas $A_j\mathbf{x}=\mathbf{b}$, para $j=1,2$, con

$$A_1=\begin{pmatrix} 1 & 1\\ 1 & 0\end{pmatrix},\quad A_2=\begin{pmatrix} 1 & 1\\ 1 & 0.01\end{pmatrix},\quad \mathbf{b}=\begin{pmatrix} 100\\ 1\end{pmatrix}.$$

   ¿Qué se puede decir? Calcúlese el *vector residual* $\mathbf{r}=A_j\tilde{\mathbf{x}}-\mathbf{b}$ y el vector real $\mathbf{e}=\mathbf{x}-\tilde{\mathbf{x}}$, para $j=1,2$, donde $\tilde{\mathbf{x}}$ denota la solución calculada y $\mathbf{x}$ la solución exacta.

3. *Coste operativo.* Una forma de medir la eficiencia del método de eliminación de Gauss es contar el número de operaciones aritméticas que se necesitan para resolver un sistema lineal. Por convención, se cuentan solo el número de multiplicaciones y divisiones porque la mayoría de los ordenadores realizan las sumas y restas de forma mucho más rápida. Además, el número de multiplicaciones y divisiones se cuentan conjuntamente. Demuéstrese entonces que el numero total de multiplicaciones y divisiones necesarias para obtener la solución de un sistema lineal de orden n mediante el método de eliminación de Gauss es $\frac{n^3}{3}+n^2-\frac{n}{3}$. (Así el número de operaciones aritméticas necesarias para resolver un sistema lineal por el método de eliminación de Gauss es aproximadamente $\frac{n^3}{3}$; es decir, de orden $\mathcal{O}(n^3)$.)

4. El *método de Gauss-Jordan* es una variación del método de Gauss. La principal diferencia consiste en que cuando se elimina una incógnita, se elimina de todas las ecuaciones, no solo de las subsiguientes. Entonces, el paso de elimación genera una matriz diagonal en vez de una triangular. En consecuencia, no es necesario utilizar la sustitución regresiva para obtener la solución. Ilústrese dicho método con los siguientes sistemas:

$$a)\left\{\begin{array}{l} 2x+y-z=1\\ 5x+2y+2z=-4\\ 3x+y+z=5\end{array}\right. \text{ sin utilizar pivoteo,}$$

$$b)\left\{\begin{array}{l} x+y-z=-3\\ 6x+2y+2z=2\\ -3x+4y+z=1\end{array}\right. \text{ utilizando pivoteo parcial.}$$

5. La *inversa de una matriz* cuadrada de orden n se puede calcular resolviendo los n sistemas de ecuaciones $A\mathbf{x_j}=\mathbf{e_j}$, con $j=1,2,\ldots,n$, y donde $\mathbf{e_j}$ es el vector que tiene todas las componentes ceros excepto la j-ésima que es uno, puesto que los vectores soluciones $\mathbf{x_1},\mathbf{x_2},\ldots,\mathbf{x_n}$ son respectivamente las columnas $1,2,\ldots,n$ de A^{-1}. Calcúlese, utilizando el método de eliminación de Gauss sin pivoteo, la matriz inversa de

$$A=\begin{pmatrix} 6 & 3 & 11\\ 3 & 2 & 7\\ 3 & 2 & 6\end{pmatrix}.$$

6. Muéstrese que la factorización LU de la matriz A se puede utilizar para calcular A^{-1}. (Utilícese el ejercicio anterior para calcular la matriz inversa.)

7. Dadas las matrices

$$A=\begin{pmatrix} 1 & 2 & 3 & 4\\ 1 & 4 & 9 & 16\\ 1 & 8 & 27 & 64\\ 1 & 16 & 81 & 256\end{pmatrix} \quad \text{y} \quad B=\begin{pmatrix} 1 & 1 & 0 & 0 & 0\\ 1 & 2 & 1 & 0 & 0\\ 0 & 1 & 3 & 1 & 0\\ 0 & 0 & 1 & 4 & 1\\ 0 & 0 & 0 & 1 & 5\end{pmatrix},$$

   encuéntrense sus descomposiciones LU. ¿Se puede predecir la existencia de dichas descomposiciones sin necesidad de calcularla? ¿Cómo se pueden emplear estas descomposiciones para calcular sus inversas?

8. Dada la matriz tridiagonal de orden n

$$A_n = \begin{pmatrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 2 & -1 \\ & & & & -1 & 2 \end{pmatrix},$$

determínese una fórmula general para $A_n = LU$, donde LU es la factorización de Doolittle. (*Ayuda:* estudiar los casos $n = 3, 4, 5$ y deducir una fórmula general.)

9. Sean

$$A = \begin{pmatrix} 2 & -6 & 1 \\ -1 & 7 & -1 \\ 1 & -3 & 2 \end{pmatrix} \qquad \text{y} \qquad \mathbf{b} = \begin{pmatrix} 12 \\ -8 \\ 16 \end{pmatrix}.$$

a) Encuéntrese la factorización de Crout de A.

b) Resuélvase el sistema $A\mathbf{x} = \mathbf{b}$ a partir de la factorización anterior.

10. Encuéntrense las factorizaciones de Cholesky de las matrices

$$A = \begin{pmatrix} 13 & 11 & 11 \\ 11 & 13 & 11 \\ 11 & 11 & 13 \end{pmatrix}, \quad B = \begin{pmatrix} 2.25 & -3 & 4.5 \\ -3 & 5 & -10 \\ 4.5 & -10 & 34 \end{pmatrix}, \quad C = \begin{pmatrix} 15 & -18 & 15 & -3 \\ -18 & 24 & -18 & 4 \\ 15 & -18 & 18 & -3 \\ -3 & 4 & -3 & 1 \end{pmatrix}.$$

11. Sea la matriz tridiagonal de orden n

$$A_n = \begin{pmatrix} a & 1 & & & & \\ 1 & a & 1 & & & \\ & 1 & a & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & a & 1 \\ & & & & 1 & a \end{pmatrix}.$$

a) Demuéstrese que A_n es simétrica definida positiva para $a \geq 2$.

b) Si $a \geq 2$, encuéntrese un método recurrente para hallar la factorización de Cholesky.

12. Discútase la convergencia de los métodos de Jacobi y Gauss-Seidel cuando se aplican para resolver los siguientes sistemas lineales:

$$a)\ \begin{cases} 2x + 3z = 1 \\ 4y + 2z = 2 \\ y + 6z = 3 \end{cases} \qquad b)\ \begin{cases} x + 2y - 2z = 1 \\ x + y + z = 1 \\ 2x + 2y + z = 1 \end{cases} \qquad c)\ \begin{cases} x + y + 2z = 1 \\ x + 3y + z = 1 \\ y + 4z = 1 \end{cases}$$

$$d)\ \begin{cases} 2x - y + z = 1 \\ 2x + 2y + 2z = 1 \\ -x - y + 2z = 1 \end{cases} \qquad e)\ \begin{cases} 3x + y + z = 4 \\ -2x + 4y = 1 \\ -x + 2y - 6z = 2 \end{cases} \qquad f)\ \begin{cases} 8x + 2y + 3z = 51 \\ 2x + 5y + z = 23 \\ -3x + y + 6z = 20 \end{cases}$$

A continuación, calcúlense, partiendo del vector cero, las soluciones aproximadas después de realizar tres iteraciones.

13. Estúdiese la convergencia de los métodos de Jacobi y Gauss-Seidel para resolver un sistema lineal cuya matriz de coeficientes es

$$A = \begin{pmatrix} a & 0 & 1 \\ 0 & a & 0 \\ 1 & 0 & a \end{pmatrix}, \quad a \in \mathbb{R}.$$

Cuando ambos métodos converjan a la vez, justifíquese cuál lo hace más rápido.

14. Si tenemos que resolver un sistema lineal en el que la matriz de coeficientes es diagonal estrictamente dominate y tridiagonal ¿convergen los métodos de Jacobi y Gauss-Seidel? ¿cuál lo hará de forma más rápida? Razónense las respuestas.

15. Descríbase el valor de cada componente de la iteración $\mathbf{x}^{(k)}$ del método SOR y dése la forma explícita de la correspondiente matriz de iteración para cada $k \in \mathbb{N}$. Ilústrese la primera iteración del método, utilizando $\omega = 1.2$ y empezando con el vector $\mathbf{x}^{(0)} = (0,0,0)^T$, para resolver el sistema lineal $A\mathbf{x} = \mathbf{b}$, donde

$$A = \begin{pmatrix} 4 & -2 & 0 \\ -2 & 6 & -5 \\ 0 & -5 & 11 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 8 \\ -29 \\ 43 \end{pmatrix}.$$

16. El *condicionamiento de una matriz* A se define mediante el número de condición $\kappa(A) = \|A\|\|A^{-1}\|$, donde $\|\cdot\|$ es una norma matricial, y cumple que $\kappa(A) \geq 1$. En general, $\kappa(A)$ depende de la elección de la norma. Si $\kappa(A)$ es grande, pequeños cambios en el sistema provocan grandes cambios en la solución, diciéndose entonces que el sistema está mal condicionado. Sean los sistemas $A\mathbf{x} = \mathbf{b_j}$ con $i = 1, 2$ y

$$A = \varepsilon \begin{pmatrix} \varepsilon & 1 & 0 \\ -\varepsilon & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad \mathbf{b_1} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \qquad \mathbf{b_2} = \begin{pmatrix} 0.9 \\ 1.1 \\ 1 \end{pmatrix}, \qquad \varepsilon = 10^{-7}.$$

a) Calcúlese $\kappa(A)$ con las normas 1 e infinito.

b) Resuélvanse los sistemas y coméntense los resultados.

Capítulo 3

Resolución de ecuaciones no lineales

3.1. Introducción

Estudiamos en este capítulo uno de los problemas más básicos de la aproximación numérica y con mayor historia: el cálculo de raíces. Es decir, la determinación de una *raíz*, o solución, de una ecuación de la forma $f(\mathbf{x}) = 0$, donde f, por lo general, es una función real no lineal de variables reales. Las raíces de esta ecuación también se llaman *ceros* de la función f. Un ejemplo que muestra la necesidad de tener técnicas de aproximación a alguna solución del problema anterior es el caso simple de encontrar soluciones de un polinomio. Se conocen fórmulas para polinomios de grados 2, 3 y 4, siendo de complejidad creciente, pero con la posibilidad de determinar sus ceros exactamente. A principios de siglo XIX Galois probó que no existen fórmulas explícitas para determinar los ceros de polinomios de grado mayor o igual que 5. La situación es aún más difícil cuando f no es un polinomio. Esta limitación obliga a buscar métodos para encontrar los ceros de forma aproximada. Los métodos que se discuten en esta lección son iterativos y de dos tipos: uno en el que se puede asegurar la convergencia y el otro en el que la convergencia depende de una o dos aproximaciones iniciales.

Métodos conceptualmete sencillos son los métodos de intervalo, que aprovechan el cambio de signo de la función en el entorno de una raíz. Así, partiendo de dos valores proporcionados inicialmente, dichos métodos tratan de reducir el tamaño del intervalo que encierra al cero buscado, hasta converger a éste con la suficiente precisión. Un ejemplo de este tipo de métodos es el de bisección, que además de ser un método simple e intuitivo, se puede utilizar para obtener una adecuada estimación inicial del cero de la función que después puede ser refinado por métodos más poderosos.

El segundo tipo de métodos se basa en aproximar la función, cuyos ceros se buscan, por una recta. Los métodos que describiremos parten de una aproximación (método de Newton, que aproxima la función por una recta tangente) o dos aproximaciones (método de la secante, que aproxima la función por una recta secante) y determinan el cero de la función con una precisión deseada. Estos métodos no garantizan la convergencia, pero, cuando convergen, lo hacen generalmente más rápidamente que los métodos de intervalo ([5]).

3.2. El método de bisección

La primera y más elemental técnica que consideramos es el *método de bisección*, que está basado en una propiedad bien conocida de las funciones continuas: el *teorema del valor medio*. Este método se emplea para determinar una raíz de $f(x) = 0$ en un intervalo $[a, b]$, supuesto que f es continua en dicho intervalo y que $f(a)$ y $f(b)$ tienen signos distintos. Aunque el método funciona en el caso en que haya más de una raíz en el intervalo $[a, b]$, supondremos por simplicidad que la raíz es única y la llamaremos α. (En el caso de varias raíces, habría que localizar un intervalo que contenga sólo una de ellas.)

El método de bisección emplea la idea anterior de la siguiente manera: si $f(a)f(b) < 0$, entonces calculamos $c = (a + b)/2$ y averiguamos si $f(a)f(c) < 0$. Si lo es, entonces f tiene un cero en $[a, c]$. A continuación renombramos c como b y comenzamos otra vez con el nuevo intervalo $[a, b]$, cuya longitud es igual a la mitad del intervalo original. Si $f(a)f(c) > 0$, entonces $f(c)f(b) < 0$, y en este caso renombramos a c como a. En ambos casos se ha generado un nuevo intervalo que contiene un cero de f, y el proceso de partir por la mitad

el nuevo intervalo puede repetirse hasta que la raíz se localiza con tanta exactitud como se quiera, es decir,

$$|a_n - b_n| < \epsilon,$$

donde a_n y b_n son los extremos del n-ésimo intervalo $[a_n, b_n]$ y ϵ es una tolerancia especificada. El método de bisección también se conoce como *método de bipartición*. Véase la figura 3.1.

Algunos otros criterios de parada diferentes del anterior que pueden utilizarse son:

$$\frac{|a_n - b_n|}{|a_n|} < \epsilon \qquad \text{o} \qquad |f(a_n)| < \epsilon.$$

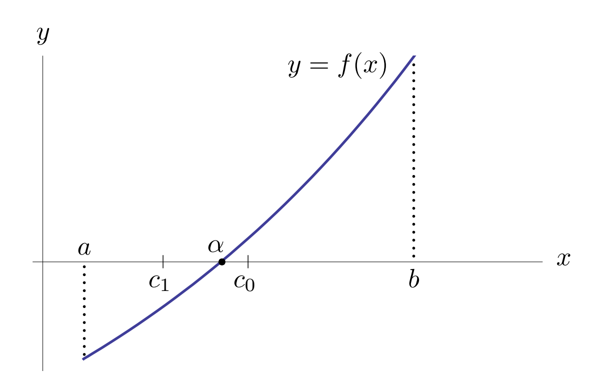


Figura 3.1: El método de bisección y las dos primeras aproximaciones a la raíz α.

EJEMPLO. La función $f(x) = x^3 - x^2 - 1$ tiene exactamente un cero en $[1, 2]$. Vamos a aproximarlo mediante el método de bisección. Como $f(1) = -1 < 0$ y $f(2) = 3 > 0$, tenemos que $f(1)f(2) < 0$. Empezamos entonces con $a_0 = 1$ y $b_0 = 2$ para calcular

$$c_0 = \frac{a_0 + b_0}{2} = \frac{1+2}{2} = 1.5 \qquad \text{y} \qquad f(c_0) = f(1.5) = 0.125.$$

Ahora $f(1)f(1.5) < 0$, luego la función cambia de signo en $[a_0, c_0] = [1, 1.5]$. Para continuar, tomamos entonces $a_1 = a_0$ y $b_1 = c_0$, de manera que

$$c_1 = \frac{a_1 + b_1}{2} = \frac{1+1.5}{2} = 1.25 \qquad \text{y} \qquad f(c_1) = f(1.25) = -0.609375.$$

De nuevo $f(1.25)f(1.5) < 0$, luego la función cambia de signo en $[1.25, 1.5]$. Elegimos ahora $a_2 = c_1$ y $b_2 = b_1$. Continuando de esta manera, se puede ver en [16] que hay convergencia a la raíz $\alpha = 1.465454$ después de doce iteraciones con una tolerancia de 10^{-4}. □

Ya hemos indicado anteriormente que es importante, para un método que proporciona una sucesión de aproximaciones, saber si la sucesión converge y en cómo de rápido lo hace en dicho caso.

Para una sucesión convergente, dos cantidades describen la rapidez con que la sucesión lo hace: el *orden de convergencia* y la *razón de convergencia*. Se suele decir que la convergencia es lineal si el orden de convergencia es uno y cuadrática si es dos.

Una forma de definir la convergencia lineal, que es particularmente conveniente para el método de bisección, es que la razón de error en el paso n-ésimo dividido por el error en el paso $(n-1)$-ésimo se aproxima a un valor constante (la razón de convergencia) cuando $n \to \infty$. Aunque no sepamos que el error se reduce en $1/2$ en cada paso del método de bisección, la cota del error se reduce en $1/2$ y el límite de la razón de error también se aproxima a $1/2$. Entonces, el método de bisección converge linealmente con razón $1/2$. Damos a continuación un resultado sobre la **convergencia** del método,que pone de manifiesto lo que acabamos de decir:

> Si $[a_0, b_0], [a_1, b_1], \ldots, [a_n, b_n], \ldots$ denotan los intervalos en el método de bisección, entonces los límites $\lim_n a_n$ y $\lim_n b_n$ existen, son iguales y representan un cero de f. Si $\alpha = \lim_n c_n$ y $c_n = (a_n + b_n)/2$, entonces $|\alpha - c_n| = (b_0 - a_0)/2^{n+1}$.

El método de bisección siempre converge, suponiendo que el intervalo con el que se comienza dicho método contenga una raíz, y es fácil de programar, comparado con otros métodos, aunque su velocidad de convergencia sea lenta por ser lineal.

COMENTARIO ADICIONAL. ▷ El método de bisección es útil para dar una idea rápida de la *localización* de las raíces, es decir, para determinar intervalos de longitud pequeña que contengan una única raíz; en cambio, al ser de convergencia muy lenta, los cálculos aumentan considerablemente si deseamos una buena aproximación de la raíz. Esto hace que este método se aplique, principalemente, como un paso previo a la utilización de otros métodos iterativos de convergencia más rápida.

3.3. El método de Newton

El método de Newton es un procedimiento general que se puede aplicar en muy diversas situaciones. Cuando se emplea para localizar los ceros de una función real de variable real se suele llamar *método de Newton-Raphson*.

La idea del método de Newton para resolver la ecuación $f(x) = 0$ es aproximar la función f por su tangente l en una aproximación de la raíz α y resolver la ecuación $l(x) = 0$ (lo que se denomina *linealización* de la función). Tomamos esta solución como nueva estimación de la raíz de $f(x) = 0$ y repetimos el proceso hasta obtener la precisión deseada.

Sea x_0 la estimación inicial de la solución. Hallamos la tangente a la gráfica de f en el punto $(x_0, f(x_0))$ y tomamos la intersección x_1 de la tangente con el eje de las x como nueva estimación de la solución. Así, la ecuación de la recta tangente es

$$y = f(x_0) + f'(x_0)(x - x_0)$$

y la intersección con el eje de las x se obtiene haciendo $y = 0$:

$$x_1 \equiv x = x_0 - \frac{f(x_0)}{f'(x_0)}$$

siempre que la derivada f' no se anule en x_0.

Procediendo de forma iterativa, véase la figura 3.2, en el paso n-ésimo determinamos:

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}, \quad n = 0, 1, 2 \ldots$$

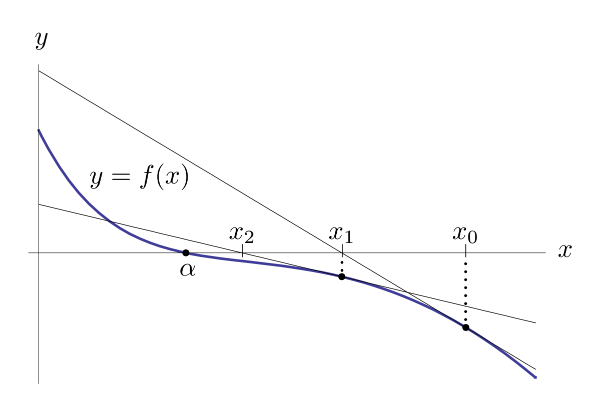


Figura 3.2: El método de Newton y las dos primeras aproximaciones a la raíz α.

En teoría, el método de Newton converge a la raíz α sólo después de un número infinito de iteraciones. En la práctica, se requiere una aproximación de α hasta una tolerancia especificada ϵ, de modo que las iteraciones pueden terminarse si

$$|f(x_n)| \leq \epsilon, \qquad |x_{n+1} - x_n| \leq \epsilon \qquad \text{o} \qquad \frac{|x_{n+1} - x_n|}{|x_{n+1}|} \leq \epsilon.$$

EJEMPLO. Utilicemos el método de Newton para aproximar una raíz de $x^3-x^2-1=0$ empezando en $x_0=1$. Como $f'(x)=3x^2-2x$, tenemos que las dos primeras iteraciones de Newton son

$$x_1 = x_0 - f(x_0)/f'(x_0) = 1-(-1)/1 = 2, \qquad x_2 = x_1 - f(x_1)/f'(x_1) = 2-3/8 = 1.625.$$

Continuando de esta manera, se llega, véase [16], a que hay convergencia a la raíz $\alpha = 1.465571$ después de seis iteraciones con una tolerancia de 10^{-4}. □

El método de Newton suele proporcionar resultados precisos en unas pocas iteraciones. Con la ayuda del polinomio de Taylor de f en α de primer orden podremos ver por qué. Supongamos que α es la solución de $f(x)=0$ y que f'' existe en un intervalo en el que están tanto α como x_n. Evaluando $x=\alpha$ la fórmula correspondiente al primer polinomio de Taylor de f en x_n obtenemos:

$$0 = f(\alpha) = f(x_n) + f'(x_n)(\alpha - x_n) + \frac{1}{2} f''(\xi)(\alpha - x_n)^2,$$

donde ξ está entre x_n y α. En consecuencia, si $f'(x_n)\neq 0$, tenemos

$$\alpha - x_n + \frac{f(x_n)}{f'(x_n)} = -\frac{f''(\xi)}{2f'(x_n)}(\alpha - x_n)^2.$$

Y como $x_{n+1} = x_n - \dfrac{f(x_n)}{f'(x_n)}$, podemos escribir

$$\alpha - x_{n+1} = -\frac{f''(\xi)}{2f'(x_n)}(\alpha - x_n)^2.$$

Si $f'(x)\neq 0$ en un intervalo I alrededor de α y x_n está en I, entonces

$$|\alpha - x_{n+1}| \leq \frac{|f''(\xi)|}{2\,|f'(x_n)|}|\alpha - x_n|^2 \leq K(\alpha - x_n)^2, \quad \text{donde} \quad K = \frac{1}{2}\,\frac{\max\limits_{x\in I}|f''(x)|}{\min\limits_{x\in I}|f'(x)|}.$$

El rasgo destacable de esta desigualdad es que el error $|\alpha - x_{n+1}|$ cometido en la aproximación $(n+1)$-ésima está acotado por, más o menos, el cuadrado del error $|\alpha - x_n|$ cometido en la aproximación n-ésima. En la práctica, esto quiere decir que el número de cifras precisas en el método de Newton tiende, aproximadamente, a duplicarse tras cada iteración, lo que se denomina *convergencia cuadrática* u orden de convergencia dos.

La bondad del método de Newton se basa en la hipótesis de que la derivada de f no se anule en las aproximaciones a la raíz α. Si f' es continua, esto significa que la técnica será satisfactoria siempre que $f'(\alpha)\neq 0$ y se use una aproximación inicial lo suficientemente precisa. La condición $f'(\alpha)\neq 0$ no es banal; se verifica precisamente cuando α es una raíz simple. Cuando la raíz no es simple, el método de Newton puede ser convergente, pero no con la velocidad indicada anteriormente.

El método de Newton es una técnica poderosa, pero tiene algunas dificultades. Una de ellas es que si el aproximación inicial x_0 no está suficientemente cerca de la raíz α, el método puede no converger. Tampoco se debería elegir x_0 tal que $f'(x_0)$ sea próximo a cero, puesto que en este caso la tangente es casi horizontal. La convergencia del método de Newton a la raíz α dependerá, en general, de la elección del aproximación inicial x_0.

La mayoría de las veces solo sabemos dar resultados de convergencia local, que solo son válidos para x_0 perteneciendo a un cierto entorno de la raíz α. Los métodos que convergen a α para cualquier elección de x_0 perteneciente a un intervalo se dicen globalmente convergentes a α.

El siguiente resultado de **convergencia local** para el método de Newton ilustra la importancia de la elección de la aproximación inicial.

> Sean f, f' y f'' continuas y f' no nula en algún intervalo abierto que contenga a una raíz simple α de $f(x)=0$. Entonces, existen a y b, con $a<\alpha<b$, tales que el método de Newton converge a α para cualquier aproximación inicial $x_0\in(a,b)$.

Hay situaciones particulares en las que se tiene la seguridad de que el método de Newton converge a partir de cualquier aproximación inicial que se elija (**convergencia global**). Como muestra, véase [24], damos el siguiente resultado sobre las condiciones suficientes de convergencia del método de Newton:

Sea f dos veces derivable con continuidad en un intervalo $[a, b]$ y tal que:

a) $f(a)f(b) < 0$,
b) $f'(x) \neq 0, \ \forall x \in [a, b]$,
c) $f''(x)$ no cambia de signo en $[a, b]$.

Entonces, existe una única raíz α de $f(x) = 0$ en $[a, b]$ y el método de Newton converge a α para todo aproximación inicial $x_0 \in [a, b]$ cumpliendo que:

i) $f(x_0)f''(x_0) \geq 0$, o bien,
ii) $f(x_0)f''(x_0) < 0$ y $x_1 \in [a, b]$.

Comentario adicional. ▷ Cuando el método de Newton converge, suele hacerlo de forma rápida. Cuando no converge, suele ser porque la aproximación inicial no está suficientemente cerca de la solución. Problemas típicos de convergencia suelen aparecer cuando el valor de $f'(x)$ está próximo a cero en un entorno de la solución, donde $f(x) = 0$.

3.4. El método de la secante

El método de Newton converge muy rápidamente, pero necesita evaluar la derivada de la función en cada paso y es muy sensible a la estimación inicial. Para superar esta dificultad se puede reemplazar $f'(x_n)$ en la expresión del método de Newton por el cociente de diferencias:

$$f'(x_n) \approx \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}}.$$

Esta aproximación tiene su origen en la definición de f' en términos de un límite:

$$f'(x) = \lim_{u \to x} \frac{f(x) - f(u)}{x - u}.$$

Cuando se realiza esta sustitución, el algoritmo resultante se conoce como *método de la secante* y su expresión es:

$$\text{dados } x_{-1}, x_0; \quad x_{n+1} = x_n - f(x_n)\left(\frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})}\right), \quad n = 0, 1, 2 \dots$$

Ya que el cálculo de x_{n+1} requiere conocer x_n y x_{n-1}, se deben dar al principio los valores de dos aproximaciones iniciales (véase la figura 3.3). Sin embargo, cada x_{n+1} requiere sólo una nueva evaluación de f.

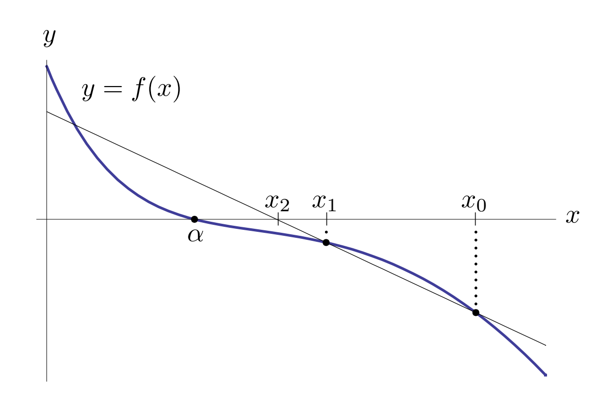


Figura 3.3: El método de la secante y la primera aproximación a la raíz α.

Un criterio adecuado de parada es

$$|f(x_n)| \leq \epsilon, \qquad |x_{n+1} - x_n| \leq \epsilon \qquad \text{o} \qquad \frac{|x_{n+1} - x_n|}{|x_{n+1}|} \leq \epsilon,$$

donde ϵ es una tolerancia especificada.

Ejemplo. Utilizamos ahora el método de la secante con $x_0 = 1$ y $x_1 = 2$ para aproximar la raíz real de $x^3 - x^2 - 1 = 0$. Como $f(1) = -1$ y $f(2) = 3$, la primera aproximación a la raíz es:

$$x_2 = x_1 - f(x_1)\left(\frac{x_1 - x_0}{f(x_1) - f(x_0)}\right) = 2 - 3\,\frac{2-1}{3-(-1)} = 1.25.$$

Ahora $f(x_2) = f(1.25) = -0.609375$ y la siguiente aproximación es:

$$x_3 = x_2 - f(x_2)\left(\frac{x_2 - x_1}{f(x_2) - f(x_1)}\right) = 1.25 - (-0.609375)\frac{1.25-2}{-0.609375-3} = 1.376623.$$

Continuando de esta manera, se llega, véase [16], a que hay convergencia a la raíz $\alpha = 1.465571$ después de siete iteraciones con una tolerancia de 10^{-4}. □

La interpretación geométrica del método de la secante es similar a la del método de Newton, puesto que ahora la recta tangente a la curva se reemplaza por una recta secante.

El método de la secante tiene las ventajas de no utilizar la derivada y ser más robusto que el de Newton. En cambio, es más lento que este último, pero más rápido que el de bisección, lo que puede verse en el siguiente resultado de **convergencia local**:

> Sean f, f' y f'' continuas y f' no nula en algún intervalo abierto que contenga a una raíz simple α de $f(x) = 0$. Entonces, existen a y b, con $a < \alpha < b$, tales que, para cualesquier aproximaciones iniciales $x_{-1}, x_0 \in (a,b)$, el método de la secante converge a α con orden de convergencia $\frac{1}{2}\left(1+\sqrt{5}\right) \approx 1.618$.

3.5. El método de Newton para sistemas de ecuaciones no lineales

El método de Newton para sistemas de ecuaciones no lineales es una extensión del método utilizado para resolver una sola ecuación, por tanto sigue la misma estrategia que se empleó en el caso de una sola ecuación: linealizar y resolver, repitiendo los pasos con la frecuencia necesaria. Consideramos, por sencillez, el caso de dos ecuaciones con dos variables:

$$\begin{cases} f(x,y) = 0, \\ g(x,y) = 0. \end{cases}$$

Si definimos la función $F(x,y) = (f(x,y), g(x,y))$, transformamos el sistema anterior en

$$F(x,y) = 0.$$

Podemos realizar entonces una formulación idéntica al caso de una variable donde la derivada está dada en este caso por la matriz jacobiana

$$F'(x,y) = \begin{pmatrix} \dfrac{\partial f}{\partial x}(x,y) & \dfrac{\partial f}{\partial y}(x,y) \\ \dfrac{\partial g}{\partial x}(x,y) & \dfrac{\partial g}{\partial y}(x,y) \end{pmatrix}.$$

Así, el método de Newton comienza por un vector inicial $(x_0, y_0)^T$ y calcula el resto de aproximaciones mediante

$$\begin{pmatrix} x_{n+1} \\ y_{n+1} \end{pmatrix} = \begin{pmatrix} x_n \\ y_n \end{pmatrix} - (F'(x_n,y_n))^{-1} F(x_n,y_n),$$

donde $(F'(x_n,y_n))^{-1}$ es la matriz inversa de $F'(x_n,y_n)$. Para poder aplicar este método es necesario que $F'(x,y)$ sea no singular.

Un problema del método anterior es que el cálculo de la matriz inversa es costoso computacionalmente y debemos calcularla en cada paso. Esto se puede resolver descomponiendo el método en dos etapas:

1. Resolver el sistema lineal con dos ecuaciones y dos incognitas:

$$F'(x_n,y_n)\begin{pmatrix} u_n \\ v_n \end{pmatrix} = -F(x_n,y_n).$$

2. Tomar como nueva aproximación:

$$\begin{pmatrix} x_{n+1} \\ y_{n+1} \end{pmatrix} = \begin{pmatrix} x_n \\ y_n \end{pmatrix} + \begin{pmatrix} u_n \\ v_n \end{pmatrix}.$$

El sistema lineal de la etapa 1 se puede resolver por cualquiera de los métodos vistos en el capítulo dedicado a la resolución de sistemas lineales, y puede resultar difícil de resolver cuando la matriz $F'(x, y)$ es casi singular.

Ejemplo. Apliquemos el método de Newton para aproximar una solución del sistema no lineal dado por

$$\begin{cases} f(x, y) = x^3 + 3y^2 - 21 = 0 \\ g(x, y) = x^2 + 2y + 2 = 0 \end{cases}$$

utilizando la aproximación inicial $(x_0, y_0)^T = (1, -1)^T$. La matriz jacobiana es

$$F'(x, y) = \begin{pmatrix} 3x^2 & 6y \\ 2x & 2 \end{pmatrix}.$$

En el punto $(1, -1)$ el vector función y la matriz jacobiana tienen los valores

$$F(1, -1) = \begin{pmatrix} -17 \\ 1 \end{pmatrix} \qquad \text{y} \qquad F'(1, -1) = \begin{pmatrix} 3 & -6 \\ 2 & 2 \end{pmatrix}.$$

Hacemos ahora la primera etapa:

$$\begin{pmatrix} 3 & -6 \\ 2 & 2 \end{pmatrix} \begin{pmatrix} u_0 \\ v_0 \end{pmatrix} = -\begin{pmatrix} -17 \\ 1 \end{pmatrix} \qquad \Rightarrow \qquad \begin{pmatrix} u_0 \\ v_0 \end{pmatrix} = \begin{pmatrix} 1.555556 \\ -2.055560 \end{pmatrix}.$$

La nueva aproximación es entonces (segunda etapa):

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = \begin{pmatrix} x_0 \\ y_0 \end{pmatrix} + \begin{pmatrix} u_0 \\ v_0 \end{pmatrix} = \begin{pmatrix} 1 \\ -1 \end{pmatrix} + \begin{pmatrix} 1.555556 \\ -2.055560 \end{pmatrix} = \begin{pmatrix} 2.555556 \\ -3.055560 \end{pmatrix}.$$

Iteramos el proceso hasta que $\|(u_n, v_n)\|_\infty = \text{máx}\{|u_n|, |v_n|\} < 10^{-6}$, de manera que en la sexta iteración obtenemos la solución aproximada $(1.643038, -2.349787)^T$. Véase [16]. □

Como en el método de Newton para una ecuación, la convergencia no está garantizada. Probablemente, el método de Newton para sistemas de ecuaciones no lineales convergerá si se dan las siguientes condiciones ([12]):

a) Las funciones f y g y sus derivadas deben ser continuas y acotadas cerca de la solución,

b) el determinante de la matriz jacobiana, llamado jacobiano, debe ser distinto de cero cerca de la solución,

c) las aproximaciones iniciales de la solución deben estar suficientemente cerca de la solución.

Cuando el método de Newton converge, suele hacerlo de forma rápida. Cuando no converge, suele ser porque las aproximaciones iniciales no están suficientemente cerca de la solución. Problemas típicos de convergencia suelen aparecer cuando el valor del jacobiano está próximo a cero en un entorno de la solución, donde $F(x, y) = 0$.

Nota. El método de Newton es fácilmente generalizable al caso de sistemas de m ecuaciones no lineales con $m > 2$. Un desarrollo del mismo puede consultarse en [12].

3.6. Sugerencias para seguir leyendo

Para información adicional sobre las técnicas del cálculo de raíces se recomienda Atkinson (1989) y Ralston y Rabinowitz (2001). Para sistemas de ecuaciones no lineales pueden consultarse [15], Ortega y Reinboldt (2000) y Stoer y Bulirsch (2002).

3.7. Ejercicios

1. *a*) Si f es una función continua en un intervalo $[a, b]$ en el que $f(a)f(b) < 0$, demuéstrese que el método de bisección aproxima un cero de f con un error en el paso n-ésimo de a lo sumo $(b - a)/2^{n+1}$.

   b) Determíniese entonces el número de iteraciones que requiere el método de bisección para encontrar el único cero de $f(x) = x^3 - x^2 - 1$ en el intervalo $[1, 2]$ con un error absoluto de no más de 10^{-6}.

2. Supongamos que el método de bisección se inicia en el intervalo $[50, 63]$. ¿Cuántos pasos deben darse para calcular una raíz con una precisión de 10^{-12}?

3. Para $f(x) = 2x^3 - x^2 + x - 1$ se tiene que $f(-4)f(4) < 0$ y $f(0)f(1) < 0$, así que el método de bisección se puede aplicar tanto en $[-4, 4]$ como en $[0, 1]$. Pero ¿en cuál de los dos intervalos interesa más empezar el método? ¿Por qué? Justifíquense las respuestas.

4. *El método de la falsa posición (regula falsi).* Un inconveniente del método de bisección es que al dividir los intervalos $[a, b]$ en mitades iguales no se tienen en cuenta las magnitudes de $f(a)$ y $f(b)$. Por ejemplo, si $f(a)$ está mucho más cerca del cero que $f(b)$, parece lógico que la raíz de la ecuación $f(x) = 0$ esté más cerca de a que de b. Un método alternativo que aprovecha esta mejora es el método de la falsa posición y consiste en unir $f(a)$ y $f(b)$ mediante una recta, de manera que la intersección de esta recta con el eje de las x suele dar una mejor aproximación de la raíz. (El nombre de falsa posición (en latín, regula falsi) viene del hecho de que al reemplazar la función por una recta, obtenemos una «falsa posición» de la raíz.)

   a) Dése la fórmula de la aproximación de la raíz que se obtiene en cada paso por este método.

   b) Aproxímese, mediante este método y una tolerancia de 10^{-4}, la única raíz de $x^3 - x^2 - 1 = 0$ que está en el intervalo $[1, 2]$.

5. *a*) Para calcular el inverso de un número a, sin tener que realizar divisiones o exponenciaciones, podemos hallar un cero de la función $f(x) = \frac{1}{x} - a$ mediante el método de Newton. Determínense, para $a > 0$, aproximaciones iniciales para las que el método converge.

   b) Indíquese qué función se utilizaría en el método de Newton para calcular $\sqrt[3]{a}$, con $a > 0$, y determínense aproximaciones iniciales para las que el método converge. Repítase lo mismo para $\sqrt[4]{a}$.

6. El polinomio $p(x) = x^3 + 94x^2 - 389x + 294$ se anula para $x = 1$, $x = 3$ y $x = -98$. El punto $x_0 = 2$ parece un buena aproximación inicial para aproximar cualquiera de los dos primeros ceros del polinomio mediante el método de Newton. Aplíquese el método de Newton a partir de $x_0 = 2$. ¿Qué sucede a partir de la segunda iteración? Trátese de determinar tres aproximaciones iniciales que lleven a cada una de las soluciones.

7. Para determinar la posición de un astro o un satélite en el espacio es necesario resolver la *ecuación de Kepler*, que está dada por $M = E - e \operatorname{sen} E$, donde M y e son constantes conocidas. Tomando $M = 2$, determínese si converge el método de Newton en los siguientes casos:

   a) $e = 1/2$ y $x_0 \in (\alpha, \pi)$, donde α es la raíz de la ecuación de Kepler,

   b) $e = 1/2$ y $x_0 = \pi/2$,

   c) $e = 2$ y $x_0 \in (\alpha, \pi)$, donde α es la raíz de la ecuación de Kepler,

   d) $e = 2$ y $x_0 = 3\pi/4$,

   e) $e = 2$ y $x_0 = \pi/2$.

8. Sea la ecuación $x^3 - 3x^2 + x + 3 = 0$.

   a) Demuéstrese que la ecuación tiene una única solución en el intervalo $[-1, 0]$.

   b) Calcúlense tres iteraciones del método de Newton utilizando como aproximación inicial $x_0 = 1$. ¿Qué se puede decir?

   c) Aproxímese, mediante tres iteraciones del método de Newton, la única solución de la ecuación en el intervalo $[-1, 0]$, utilizando una aproximación inicial a partir de la cual se garantice la convergencia del método de Newton a la solución. Justifíquese la elección de la aproximación inicial.

9. *Raíces complejas de ecuaciones.* También podemos utilizar el método de Newton para calcular las raíces complejas de una ecuación $f(z) = 0$, donde $z = x + iy$ es un número complejo, mediante

$$z_{n+1} = z_n - \frac{f(z_n)}{f'(z_n)}, \quad n = 0, 1, 2 \dots$$

Para ello debemos dar un número complejo como aproximación inicial z_0 para el método de Newton. El correspondiente error se puede estimar a partir de $E = \sqrt{\mathrm{Re}(w)^2 + \mathrm{Im}(w)^2}$, donde $w = \dfrac{|z_n - z_{n-1}|}{|z_n|}$ y $\mathrm{Re}(w)$ e $\mathrm{Im}(w)$ son respectivamente las partes real e imaginaria de w. Utilícese el método de Newton para encontrar una raíz compleja de la ecuación $f(z) = z^2 - z + 4 = 0$ y calcúlese el error.

10. *El método de Newton modificado.* El método de Newton pierde su convergencia cuadrática en el caso de raíces repetidas, pasando a tener solo convergencia lineal. Sin embargo, si se conoce la multiplicidad de la raíz (o se puede estimar), es posible, para preservar la velocidad de convergencia, modificar el método de la siguiente forma:

$$x_{n+1} = x_n - m\frac{f(x_n)}{f'(x_n)}, \quad n = 0, 1, 2 \dots,$$

donde m es la multiplicidad de la raíz buscada. Este método se llama método de Newton modificado. Verifíquese lo anterior, comparando los resultados, cuando se aproxima la raíz doble más cercana a 1 de la ecuación $x^3 - 2x^2 - 0.75x + 2.25 = 0$ mediante los métodos de Newton y Newton modificado.

11. *Aceleración de la convergencia.* Una aspecto interesante a la hora de aplicar un método iterativo es la posibilidad de acelerar su convergencia (siempre que sea posible). Por ejemplo, el método de Newton tiene convergencia lineal cuando aproxima raíces múltiples, salvo que modifiquemos el método convenientemente (ejercicio anterior), aunque esta modificación requiera del conocimiento previo de la multiplicidad de la raíz. El *método Δ^2 de Aitken* es una técnica que permite acelerar la convergencia de una sucesión que converge linealmente, independientemente de su origen o ámbito de aplicación. Así, si la sucesión de aproximaciones $\{x_n\}$ converge linealmente a α y definimos $y_n = x_n - \dfrac{(x_{n+1} - x_n)^2}{x_{n+2} - 2x_{n+1} + x_n}$, para todo $n \geq 0$, entonces la sucesión $\{y_n\}$ también converge a α y, en general, más rápidamente. (En [7] puede verse la dedución de la sucesión $\{y_n\}$.) Verifíquese lo anterior para la sucesión $\{x_n\}$ dada por $x_n = \cos\left(\dfrac{1}{n}\right)$, para todo $n \geq 1$.

12. *Iteración de punto fijo.* El método de Newton es un ejemplo de método iterativo que calcula una sucesión de aproximaciones mediante una fórmula del tipo $x_{n+1} = g(x_n)$, para todo $n \geq 0$. Un método así definido se llama iteración de punto fijo. Para el método de Newton la función g es $g(x) = x - \frac{f(x)}{f'(x)}$. Obsérvese que aproximar las raíces de $f(x) = 0$ es equivalente a aproximar las raíces de $g(x) = x$, que son los puntos fijos de g.

    a) Interprétese geométricamente este método.

    b) Sea $g : I \to I$, donde I es un subconjunto cerrado de $\mathbb{R}$, tal que $|g(x) - g(y)| \leq L|x - y|$ y $L < 1$. Demuéstrese que existe una única raíz de la ecuación $g(x) = x$ que se puede obtener como límite de la sucesión $\{x_{n+1} = g(x_n)\}$, para todo $n \geq 0$, a partir de cualquier aproximación inicial $x_0 \in I$.

    c) Si aplicamos la iteración de punto fijo a la función $g(x) = 2 + (x - 2)^4$ empezando en $x_0 = 2.5$, encuéntrese el intervalo I de aproximaciones iniciales para el que el método converge.

13. Si se utiliza el método de la secante para encontrar los ceros de $f(x) = x^3 - 3x^2 + 2x - 6 = 0$ con $x_0 = 1$ y $x_1 = 2$, ¿cuál es x_2?

14. Sean las ecuaciones $x^4 - x - 10 = 0$ y $x - \mathrm{e}^{-x} = 0$. Determínense las aproximaciones iniciales y utilícense para encontrar las aproximaciones a las raíces con cuatro cifras decimales mediante el método de la secante.

15. Realícense tres iteraciones del método de Newton en dos variables para aproximar las soluciones de los sistemas

$$a)\left\{\begin{array}{l} x=\operatorname{sen}(x+y) \\ y=\cos(x-y). \end{array}\right. \qquad b)\left\{\begin{array}{l} 4x^2-y^2=0 \\ 4xy^2-x=1 \end{array}\right. \qquad c)\left\{\begin{array}{l} xy^2+x^2y+x^4=3 \\ x^3y^5-2x^5y-x^2=-2 \end{array}\right.$$

partiendo de los puntos (1,1), (0,1) y (2,2) respectivamente.

16. Verifíquese que $(1,1)$ y $(-1,-1)$ son soluciones del sistema no lineal

$$\left\{\begin{array}{l} x^2+y^2-2=0 \\ xy-1=0. \end{array}\right.$$

Explíquense las dificultades que tiene el método de Newton para hallar dichas soluciones.

Capítulo 4

Aproximación de funciones y datos

4.1. Introducción

La necesidad de aproximar funciones y datos se presenta en muchas ramas de la ingeniería y las ciencias. El concepto de aproximación se basa en reemplazar una función f por otra más simple, $\widetilde{f}$, que podamos utilizar como sustituto. Como veremos en el siguiente capítulo, esta técnica se utiliza frecuentemente en integración numérica, donde, en lugar de calcular $\int_a^b f(x)\,dx$, calculamos de forma exacta $\int_a^b \widetilde{f}(x)\,dx$, donde $\widetilde{f}$ es una función fácil de integrar (por ejemplo, un polinomio). También puede ocurrir que solo se conozca la función f parcialmente por medio de valores en alguna colección finita de datos. En estos casos construiremos una función continua $\widetilde{f}$ que represente a la colección de datos. En [23] se muestran varios ejemplos.

Sabemos que una función f se puede reemplazar por un polinomio de Taylor en un intervalo dado. Computacionalmente, esta sustitución es costosa porque requiere del conocimiento de f y sus derivadas hasta el orden n (el grado del polinomio) en un punto x_0. Además, el polinomio de Taylor puede fallar para representar a f suficientemente lejos del punto x_0, como podemos ver en el siguiente ejemplo. En la figura 4.1 se compara el comportamiento de $f(x) = 1/x$ con el de su polinomio de Taylor de grado 10 construido alrededor del punto $x_0 = 1$: $P_{10}(x) = 2 - x + (x-1)^2 - (x-1)^3 + (x-1)^4 - (x-1)^5 + (x-1)^6 - (x-1)^7 + (x-1)^8 - (x-1)^9 + (x-1)^{10}$. La afinidad entre la función y su polinomio de Taylor es muy buena en un pequeño entorno de $x_0 = 1$, mientras que resulta insatisfactoria cuando $x - x_0$ se hace grande ([23]).

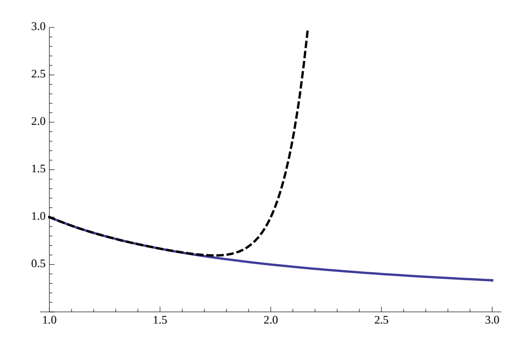


Figura 4.1: Comparación entre la función $f(x) = 1/x$ (línea continua) y su polinomio de Taylor de grado 10 alrededor del punto $x_0 = 1$ (línea discontinua).

Puesto que los polinomios de Taylor tienen la propiedad de que toda información que se utiliza está centrada en un único punto x_0, es común que estos polinomios den aproximaciones poco precisas a medida que nos vamos alejando de x_0, lo que limita la utilización de aproximaciones mediante polinomios de Taylor a situaciones en las que éstas sean necesarias únicamente en puntos cercanos a x_0. Para realizar cálculos ordinarios resulta más eficaz utilizar métodos que empleen la información disponible en varios puntos. A lo largo de este capítulo introducimos métodos de aproximación que se basan en enfoques alternativos. El uso primordial de los polinomios de Taylor en análisis numérico no es para generar aproximaciones, sino para construir técnicas numéricas.

El estudio de la teoría de aproximación implica dos tipos de problemas. Un problema surge cuando se tiene explícitamente una función pero se desea encontrar un tipo de «función más simple», como un polinomio, que se pueda usar para determinar los valores aproximados a la función dada. El otro problema consiste en ajustar funciones a datos dados y encontrar qué función, dentro de una cierta clase, es la que «mejor» representa los datos. Estudiaremos ambos problemas en este capítulo.

4.2. Interpolación

En una gran variedad de problemas es deseable representar una función a partir del conocimiento de su comportamiento en un conjunto discreto de puntos. En algunos problemas solo tendremos valores en un conjunto de datos, mientras que en otros, buscaremos representar una función mediante otra más simple. En el primer caso hablaremos de *interpolación de datos*, y en le segundo de *interpolación de funciones*. En ambos casos el objetivo es obtener estimaciones de la función en puntos intermedios, aproximar la derivada o la integral de la función en cuestión o, simplemente, obtener una representación continua o suave de las variables del problema.

La **interpolación** es el proceso de determinar una función que represente exactamente una colección de datos. El problema de interpolación se puede formular, en general, en los siguientes términos: sean (x_i, f_i), $i = 0, 1, \ldots, n$, pares de valores reales dados de manera que $x_i \neq x_j$, $i \neq j$, ¿existe alguna función $\widetilde{f}$, de tipo predeterminado, tal que $\widetilde{f}(x_i) = f_i$, para todo $i = 0, 1, \ldots, n$? Los puntos x_i se llaman *nodos* y la tal función $\widetilde{f}$ se llama *función de interpolación* del conjunto de datos $\{f_i\,;\, i = 0, 1, \ldots, n\}$. En caso de que los $\{f_i\}$ representen los valores alcanzados por una función continua f, entonces $\widetilde{f}$ se denomina función de interpolación de f.

Dependiendo del tipo de funciones de interpolación que se consideren, se distinguen varios tipos de interpolación, entre los que se pueden citar: polinómica, trigonométrica, *spline* (interpolación polinomial a trozos), racional y exponencial. El tipo más elemental de interpolación consiste en ajustar un polinomio a una colección de datos, que se conoce como *interpolación polinómica*. Los polinomios tienen derivadas e integrales que son polinomios, así que son una elección natural para aproximar derivadas e integrales. Por simplicidad, solo vamos a considerar la interpolación polinómica de Lagrange y la interpolación mediante funciones *splines*.

4.2.1. Interpolación polinómica de Lagrange

Centrémonos en la interpolación polinómica de Lagrange. Dada una tabla de $n+1$ puntos (x_i, f_i), $i = 0, 1, \ldots, n$, con $x_i \neq x_j$, $i \neq j$, llamamos interpolación polinómica a la determinación de un polinomio p_n de grado $\leq n$ y tal que

$$p_n(x_i) = f_i, \quad i = 0, 1, \ldots, n.$$

El polinomio p_n recibe el nombre de *polinomio de interpolación* de los valores f_i en los nodos x_i. Cuando f_i sea el valor de una función f en x_i, $i = 0, 1, \ldots, n$, hablaremos de interpolación polinómica de la función f en los nodos x_i, p_n se llama entonces polinomio de interpolación de f y se suele denotar por $p_n f$.

Existencia y unicidad del polinomio de interpolación

La existencia y unicidad del polinomio anterior de interpolación p_n se sigue a partir del **método de los coeficientes indeterminados** para calcular los coeficientes de p_n. Así, buscamos un polinomio p_n de grado n tal que, para $x_0 < x_1 < \cdots < x_n$ y $f_0, f_1, \ldots, f_n$, se cumpla:

$$p_n(x_i) = f_i, \quad \text{para todo} \quad i = 0, 1, \ldots, n.$$

La manera aparentemente más simple de resolver el problema es escribir:

$$p_n(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n,$$

donde a_i $(i = 0, 1, \ldots, n)$ son incógnitas a determinar mediante las relaciones:

$$f_i = p_n(x_i) = a_0 + a_1 x_i + a_2 x_i^2 + \cdots + a_n x_i^n, \quad i = 0, 1, \ldots, n.$$

Si $\mathbf{x} = (a_0, a_1, \ldots, a_n)^T$, $\mathbf{b} = (f_0, f_1, \ldots, f_n)^T$ y

$$T = \begin{pmatrix} 1 & x_0 & x_0^2 & \ldots & x_0^n \\ 1 & x_1 & x_1^2 & \ldots & x_1^n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \ldots & x_n^n \end{pmatrix},$$

las relaciones anteriores se pueden escribir como el sistema lineal

$$T\mathbf{x} = \mathbf{b}.$$

La matriz T se denomina *matriz de Vandermonde* asociada a los nodos $x_0, x_1, \ldots, x_n$ y su determinante es no nulo (por ser los puntos x_i distintos), por lo que el problema se reduce a resolver un sistema lineal de $n+1$ ecuaciones con $n+1$ incógnitas con solución única.

Ejemplo. Calculemos el polinomio que interpola la siguiente tabla de datos:

x	0	$\pi/2$
f	0	1

Como hay dos datos, el polinomio es lineal de la forma $p_1(x) = a_0 + a_1 x$. Si aplicamos la condición $p_1(x_i) = f_i$ $(i = 1, 2)$, obtenemos el sistema

$$\begin{cases} a_0 = 0 \\ a_0 + \frac{\pi}{2}\, a_1 = 1 \end{cases}$$

cuya solución es $a_0 = 0$ y $a_1 = \frac{2}{\pi}$. Por lo tanto, $p_1(x) = \frac{2}{\pi}x$. Si los datos de la tabla corresponden a la función $f(x) = \operatorname{sen} x$, vemos en la figura 4.2 la aproximación de $f(x)$ mediante el polinomio $p_1(x)$. □

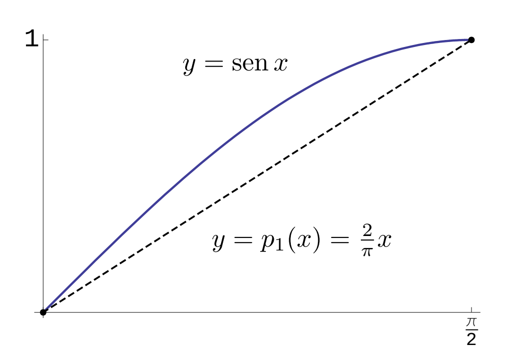


Figura 4.2: Función $y = \operatorname{sen} x$ (línea continua) y su polinomio de interpolación $y = p_1(x) = \frac{2}{\pi}x$ en los nodos $x_0 = 0$ y $x_1 = \frac{\pi}{2}$ (recta discontinua).

Comentarios adicionales.

- ▷ Es importante reseñar que si interpolamos una función f en $n+1$ puntos distintos, podemos obtener que el grado del polinomio de interpolación p_n sea distinto de n. De hecho, aunque n sea grande, el grado del polinomio puede ser pequeño. Por ejemplo, si interpolamos la función $f(x) = x^4 + 2x^3 - x^2 + x + 1$ en los puntos $\{-2, -1, 0, 1\}$, debido a la unicidad, el polinomio de interpolación es $p_3(x) = 3x + 1$.

- ▷ Aunque el método de los coeficientes indeterminados es el primer método en el que se piensa cuando se intenta hallar el polinomio de interpolación p_n, resulta excesivamente laborioso cuando n no es pequeño. Conviene entonces recurrir a otros métodos. Describimos a continuación dos alternativas que son muy adecuadas para implementarse en un ordenador: el polinomio de Lagrange y el polinomio de Newton. Aplicaremos la expresión más conveniente según sean los datos del problema.

Polinomio de interpolación de Lagrange

Comenzamos con el *polinomio de interpolación de Lagrange*, que se define de la siguiente manera:

$$p_n(x) = \sum_{i=0}^{n} f_i \ell_i(x), \quad \text{donde} \quad \ell_i(x) = \prod_{\substack{j=0 \\ j\neq i}}^{n} \frac{x - x_j}{x_i - x_j}.$$

Los polinomios ℓ_i se denominan *polinomios fundamentales de Lagrange.*

Obsérvese que $\ell_i(x_i) = 1$ y $\ell_i(x_j) = 0$, de manera que

$$p_n \text{ es un polinomio de grado } n \quad \text{y} \quad p_n(x_i) = f_i \quad (i = 0, 1, \ldots, n).$$

Luego el polinomio de interpolación de Lagrange es efectivamente el polinomio solución del problema de interpolación.

Ejemplo. Veamos cuál es el polinomio de interpolación de Lagrange que interpola los datos de la siguiente tabla:

x	0	$\pi/2$	π
f	0	1	0

Los polinomios fundamentales de Lagrange son:

$$\begin{aligned}
\ell_0(x) &= \frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)} = \frac{(x - \pi/2)(x - \pi)}{(0 - \pi/2)(0 - \pi)} = \frac{2}{\pi^2}(x - \pi/2)(x - \pi), \\
\ell_1(x) &= \frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)} = \frac{(x - 0)(x - \pi)}{(\pi/2 - 0)(\pi/2 - \pi)} = -\frac{4}{\pi^2}x(x - \pi), \\
\ell_2(x) &= \frac{(x - x_0)(x - x_1)}{(x_2 - x_0)(x_2 - x_1)} = \frac{(x - 0)(x - \pi/2)}{(\pi - 0)(\pi - \pi/2)} = \frac{2}{\pi^2}x(x - \pi/2).
\end{aligned}$$

Luego el polinomio de interpolación será:

$$p_2(x) = 0\,\frac{2}{\pi^2}(x - \pi/2)(x - \pi) - 1\,\frac{4}{\pi^2}x(x - \pi) + 0\,\frac{2}{\pi^2}x(x - \pi/2) = -\frac{4}{\pi^2}(x^2 - \pi x).$$

Si los datos de la tabla corresponden a la función $f(x) = \operatorname{sen} x$, vemos en la figura 4.3 la aproximación de $f(x)$ mediante el polinomio $p_2(x)$. □

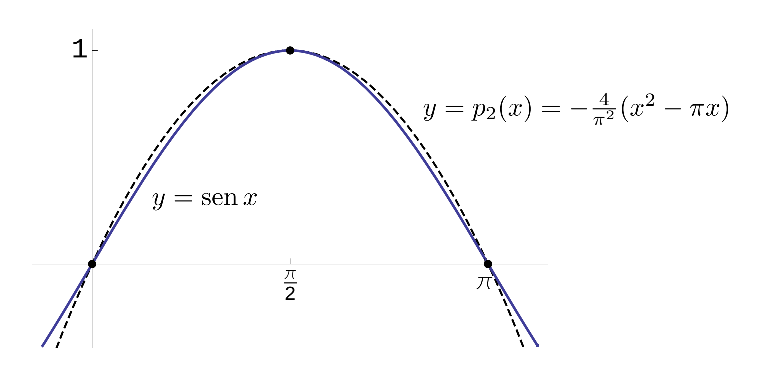


Figura 4.3: Función $y = \operatorname{sen} x$ (línea continua) y su polinomio de interpolación $y = p_2(x) = -\frac{4}{\pi^2}(x^2 - \pi x)$ en los nodos $x_0 = 0$, $x_1 = \frac{\pi}{2}$ y $x_2 = \pi$ (parábola discontinua).

Comentario adicional. ▷ Una ventaja del polinomio de interpolación de Lagrange es que su implementación en el ordenador es muy sencilla. Otra ventaja es que si queremos resolver varios problemas de interpolación con los mismos nodos, entonces solo tendremos que modificar los valores de f_i una vez calculados los polinomios fundamentales de Lagrange.

Polinomio de interpolación de Newton

Vamos a estudiar a continuación un método más eficiente de construcción del polinomio de interpolación. Si construimos el polinomio de interpolación por cualquiera de los dos métodos anteriores y, por alguna circunstancia, necesitamos añadir un nuevo punto (x_{n+1}, f_{n+1}), en ambos casos se tienen que rehacer todos los cálculos para determinar el nuevo polinomio de interpolación. La formula de Newton permite calcular este nuevo polinomio p_{n+1} conocido el polinomio anterior p_n. Para ello es necesario expresar el polinomio de interpolación p_n en términos de «diferencias» (divididas) de valores de la función en los puntos de interpolación. Así:

Dado $m \geq 0$, si definimos $f[x_i] = f_i$, $i = 0, 1, \ldots, n$, se denomina **diferencia dividida** de orden m de f en el punto x_i, y se denota por $f[x_i, x_{i+1}, \ldots, x_{i+m-1}, x_{i+m}]$, al cociente:

$$f[x_i, x_{i+1}, \ldots, x_{i+m-1}, x_{i+m}] = \frac{f[x_i, x_{i+1}, \ldots, x_{i+m-1}] - f[x_{i+1}, \ldots, x_{i+m-1}, x_{i+m}]}{x_i - x_{i+m}}.$$

A partir de las diferencias divididas, el polinomio de interpolación se determina como

$$p_n(x) = f[x_0] + \sum_{i=1}^{n} \left(f[x_0, x_1, \ldots, x_i] \prod_{j=0}^{i-1} (x - x_j) \right),$$

que se conoce como *polinomio de interpolación de Newton.*

Si añadimos a continuación un nuevo par de interpolación (x_{n+1}, f_{n+1}), el nuevo polinomio de interpolación se puede expresar como:

$$p_{n+1}(x) = p_n(x) + f[x_0, x_1, \ldots, x_{n+1}] \prod_{j=0}^{n} (x - x_j).$$

COMENTARIOS ADICIONALES. ▷ Una ventaja del polinomio de interpolación de Newton radica en el hecho de que añadir nuevos nodos no afecta a los coeficientes ya calculados, por lo que se puede ir aumentando el grado del polinomio y conseguir la precisión deseada sin tener que calcular de nuevo todos los coeficientes, ya que, por construcción, cada polinomio se obtiene del anterior mediante la última igualdad. Como consecuencia de este hecho, se suelen tomar los nodos en cierto orden, considerando primero los más próximos al punto que nos interesa, aunque el polinomio obtenido finalmente es independiente del orden considerado.

▷ Obsérvese que las diferencias divididas son recursivas, puesto que las de orden superior se calculan a partir de las de orden inferior. Esta propiedad de la recursividad es muy aprovechable a la hora de implementar eficientemente este método en un ordenador.

▷ Las expresiones de los polinomios de interpolación de Lagrange y de Newton requieren de un trabajo computacional semejante cuando solo se desea realizar una interpolación, aunque el de Lagrange es más fácil de implementar en un ordenador debido a que no requiere calcular ni almacenar las diferencias divididas. Si a priori no se conoce el grado del polinomio de interpolación, la expresión de Newton es más conveniente, como podría ocurrir por ejemplo en el caso de que fuera mejor no utilizar todos los nodos del conjunto de datos disponibles.

▷ Los polinomios de Lagrange y de Newton no proporcionan un polinomio en su forma convencional, tal y como hace el método de los coeficientes indeterminados, pero son técnicas más eficientes.

EJEMPLO. Obtengamos a continuación la fórmula de interpolación de Newton para la siguiente tabla:

x	2	4	6	8
f	4	8	14	16

Construimos el siguiente cuadro sinóptico para las diferencias divididas

$$
\begin{array}{c|lllll}
x & f & & & \\
2 & \boxed{4} & & & \\
 & & f[2,4]=\frac{8-4}{4-2}=\boxed{2} & & \\
4 & 8 & & f[2,4,6]=\frac{3-2}{6-2}=\boxed{\frac{1}{4}} & \\
 & & f[4,6]=\frac{14-8}{6-4}=3 & & f[2,4,6,8]=\frac{-1/2-1/4}{8-2}=\boxed{-\frac{1}{8}} \\
6 & 14 & & f[4,6,8]=\frac{1-3}{8-4}=-\frac{1}{2} & \\
 & & f[6,8]=\frac{16-14}{8-6}=1 & & \\
8 & 16 & & &
\end{array}
$$

Y, a partir de la diagonal superior de cuadro anterior, construimos el polinomio de interpolación: $p_3(x) = 4 + 2(x-2) + \frac{1}{4}(x-2)(x-4) - \frac{1}{8}(x-2)(x-4)(x-6) = -\frac{1}{8}x^3 + \frac{7}{4}x^2 - 5x + 8$. $\square$

Error de interpolación

A la hora de reemplazar el valor de la función f por su polinomio de interpolación $p_n f$ en puntos distintos de los puntos de interpolación es importante estimar el error:

$$E_n f(x) = f(x) - p_n f(x), \quad x \in [a,b],$$

donde $[a,b]$ es un intervalo que contiene a los nodos de interpolación.

Si la función f está tabulada y no conocemos su expresión analítica, no podemos estimar el error que se comete con el polinomio $p_n f$ cuando reemplaza a la función f.

Cuando la función f es suficientemente regular, podemos evaluar el error obtenido, cuando reemplazamos f por su polinomio de interpolación $p_n f$, mediante el siguiente resultado:

> Si $p_n f(x)$ interpola a la función $f(x)$ en los nodos distintos x_i, $i = 0, 1, \ldots, n$, y $f(x)$ tiene derivadas continuas hasta orden $n+1$ en un intervalo $[a,b]$ que contiene a los nodos de interpolación, entonces para cualquier $x \in [a,b]$, existe un $\xi \in [a,b]$ tal que
>
> $$E_n f(x) = f(x) - p_n f(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}(x-x_0)(x-x_1)\cdots(x-x_n).$$

Obviamente, $E_n f(x_i) = 0$, $i = 0, 1, \ldots, n$.

Por otra parte, si los puntos de interpolación están uniformemente distribuidos en el intervalo $[a,b]$, es decir, $x_i = a + \frac{i}{h}$, con $i = 0, 1, \ldots, n$, y $h = \frac{b-a}{n}$, entonces se puede demostrar que

$$|E_n f(x)| \leq \frac{h^{n+1}}{(n+1)!} \max_{x\in[a,b]} |f^{(n+1)}(x)|.$$

De modo que si $\max_{x\in[a,b]} |f^{(n+1)}(x)| \leq M$, donde M es una constante independiente de n, entonces el error de interpolación tiende a cero según h se va acercando a cero. Éste es el caso de las funciones trigonométricas $\operatorname{sen}(x)$, $\cos(x)$ y de la exponencial e^x en un intervalo finito.

EJEMPLO. Calculemos una cota del error cuando se aproxima $f(x) = \cos x$ mediante un polinomio de interpolación con 11 nodos igualmente espaciados en $[1,2]$. Como $n = 10$, $a = 1$, $b = 2$, $f^{(11)}(x) = \operatorname{sen} x$ y $|f^{(11)}(x)| \leq 1 = M$, se obtiene $h = 1/10$ y

$$|E_n f(x)| \leq \frac{(1/10)^{11}}{11\,!} \approx 2.50 \times 10^{-19}. \quad \square$$

La propiedad anterior (derivadas de todo orden acotadas uniformemente) no la poseen todas las funciones. El ejemplo clásico es la *función de Runge* ([5]):

Si la función $f(x) = \frac{1}{1+x^2}$ se interpola en puntos equiespaciados en el intervalo $[-5, 5]$, el error $\max_{x\in[-5,5]} |E_n f(x)|$ tiende a infinito cuando $n \to \infty$. Esto se debe al hecho de que, si $n \to \infty$, el orden de magnitud de $\max_{x\in[-5,5]} |f^{(n+1)}(x)|$ pesa más que el orden infinitesimal de $\frac{h^{n+1}}{(n+1)!}$. Obsérvese en la figura 4.4 que, al aumentar el grado del polinomio, en vez de mejorar la aproximación global, empeora, puesto que los polinomios de interpolación se desvían cada vez más de la función, sobre todo cerca de los extremos. Esta función indica que al aumentar el grado n del polinomio de interpolación, no necesariamente obtenemos una mejor reconstrucción de la misma.

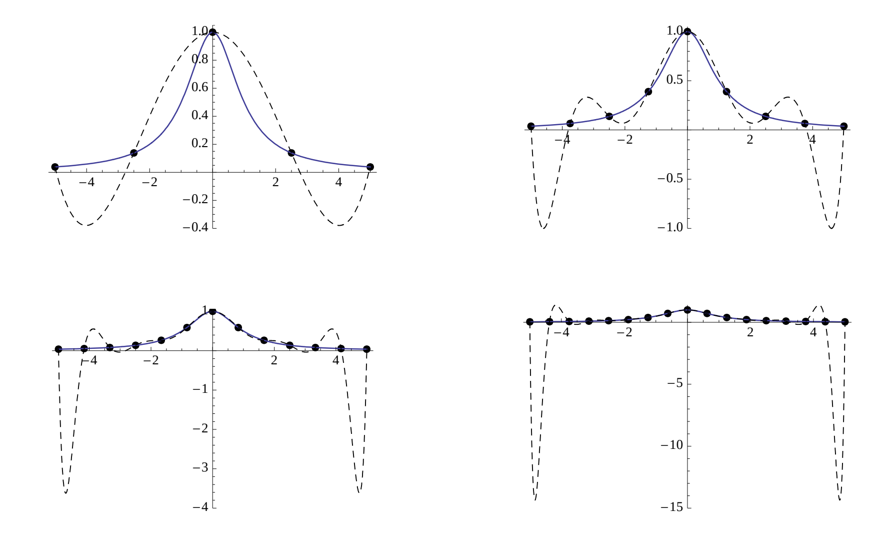


Figura 4.4: Polinomios de interpolación de grados 4, 8, 12 y 16 (líneas discontinuas) junto a la función de Runge $f(x) = \frac{1}{1+x^2}$ (línea continua).

COMENTARIO ADICIONAL. ▷ Para que la expresión anterior del error de interpolación sea útil, debemos conocer la función f, que además debe ser diferenciable, lo que puede no ocurrir. Hay una fórmula alternativa que no requiere del conocimiento previo de la función y que utiliza una diferencia dividida para aproximar $f^{(n+1)}$,

$$E_n f(x) \simeq f[x_0, x_1, \dots, x_n, x](x - x_0)(x - x_1)\cdots(x - x_n),$$

pero esta fórmula no permite encontrar el error, puesto que contiene la incógnita $f(x)$. Sin embargo, si se tiene un dato más, $f(x_{n+1})$, podemos utilizar la fórmula anterior de la siguiente forma para estimar el error:

$$E_n f(x) \simeq f[x_0, x_1, \dots, x_{n+1}](x - x_0)(x - x_1)\cdots(x - x_n).$$

Además, como el error se anula en los nodos de interpolación, deducimos la siguiente propiedad de las diferencias divididas:

> Si f es derivable con continuidad hasta la derivada n-ésima en un intervalo I y $x_0, x_1, \dots, x_{n+1}$ son $n + 1$ puntos distintos del intervalo I, entonces existe un punto ξ, comprendido entre el menor y el mayor de los x_i, tal que $f[x_0, x_1, \dots, x_n] = \dfrac{f^{(n)}(\xi)}{n!}$.

4.2.2. Interpolación mediante funciones *splines*

El problema de la interpolación polinómica de Lagrange presenta inconvenientes cuando hay muchos nodos o cuando la función a interpolar no se parece mucho a un polinomio. El aumento del número de nodos y, por tanto, del grado del polinomio, en lugar de proporcionar más precisión, provoca mayores oscilaciones del polinomio de interpolación e inestabilidad en el cálculo como hemos visto para la función de Runge.

Para evitar estos problemas, podemos utilizar una alternativa que consiste en dividir el intervalo en un número finito de subintervalos y construir un polinomio de aproximación diferente en cada subintervalo, lo que se conoce como **aproximación polinómica segmentaria** o **aproximación polinomial a trozos**. Esta

aproximación evita los inconvenientes mencionados anteriormente limitando el grado del polinomio. Como contrapartida, las condiciones de interpolación se satisfacen localmente, utilizando polinomios distintos en cada intervalo. La curva así construida se conoce como *spline*, que es un polinomio a trozos lo más suave posible en todos los puntos en los que los polinomios se juntan. La obtención de los polinomios es numéricamente estable y puede hacerse utilizando matrices dispersas (matrices con muchos ceros), véanse [5, 7].

La aproximación polinomial a trozos más habitual, por sus propiedades de suavidad y simplicidad, utiliza polinomios cúbicos en cada subintervalo y se llama interpolación mediante ***splines* cúbicos**. Un *spline* cúbico se define de la siguiente manera.

> Dados los puntos $a = x_0 < x_1 < \cdots < x_n = b$, llamamos función *spline* cúbico a una función $S(x)$ definida en $[a, b]$ tal que:
>
> - la restricción de $S(x)$ a cada subintervalo $[x_j, x_{j+1}]$, que denotaremos por $S_j(x)$, $j = 0, 1, \dots, n-1$, es un polinomio cúbico,
> - $S(x)$ es derivable con continuidad hasta la segunda derivada en $[a, b]$.

Notemos que el intervalo $[a, b]$ se subdivide en varios subintervalos y los valores que dividen el intervalo se llaman *nudos*, que pueden ser los mismos que los nodos, pero que no siempre es el caso.

Para construir el *spline* cúbico $S(x)$ que interpole a f en los nodos

$$a = x_0 < x_1 < \cdots < x_n = b,$$

aplicamos la primera condición de la definición a los polinomios cúbicos y escribimos $S(x)$ en cada subintervalo $[x_j, x_{j+1}]$ de la forma:

$$S_j(x) = a_j + b_j(x - x_j) + c_j(x - x_j)^2 + d_j(x - x_j)^3 = S(x), \quad j = 0, 1, \dots, n-1,$$

donde a_j, b_j, c_j y d_j son las constantes a determinar ($4n$ incognitas).

Como queremos que S interpole a f, se tiene que cumplir que

$$S(x_j) = f_j = f(x_j) \quad \Rightarrow \quad a_j = f(x_j), \quad j = 0, 1, \dots, n,$$

de manera que tenemos $n + 1$ ecuaciones.

Para determinar b_j, c_j y d_j, usamos la segunda condición de la definición, que dice que S, S' y S'' son continuas. Por lo tanto,

$$S_j(x_{j+1}) = S_{j+1}(x_{j+1}), \qquad S_j'(x_{j+1}) = S_{j+1}'(x_{j+1}), \qquad S_j''(x_{j+1}) = S_{j+1}''(x_{j+1}),$$

para cada $j = 0, 1, \dots, n-2$. Cada una de estas igualdades da $n - 1$ ecuaciones.

En total, se tienen $(n+1)+3(n-1) = 4n-2$ ecuaciones, pero $4n$ incognitas. Entonces, es necesario imponer dos condiciones más al *spline* cúbico S para poder determinarlo. Las dos opciones más frecuentemente usadas son:

- Condiciones de frontera libre: $S''(x_0) = 0$ ($\Rightarrow c_0 = 0$) y $S''(x_n) = 0$ ($\Rightarrow c_n = 0$); la función S resultante se llama *spline cúbico natural*.
- Condiciones de frontera fija: $S'(x_0) = f'(x_0)$ ($\Rightarrow b_0 = f'(x_0)$) y $S'(x_n) = f'(x_n)$ ($\Rightarrow b_n = f'(x_n)$); S recibe el nombre de *spline cúbico sujeto*.

En ambos casos se tienen $4n$ ecuaciones con $4n$ incognitas.

Como los términos $x_{j+1} - x_j$ aparecerán repetidamente en el desarrollo, es conveniente que introduzcamos una notación más simple:

$$h_j = x_{j+1} - x_j, \quad j = 0, 1, \dots, n-1.$$

De $S_{j+1}(x_{j+1}) = S_j(x_{j+1})$ se sigue que

$$a_{j+1} = f(x_{j+1}) = S_{j+1}(x_{j+1}) = S_j(x_{j+1}) = a_j + b_j h_j + c_j h_j^2 + d_j h_j^3, \quad j = 0, 1, \dots, n-1. \tag{4.1}$$

Como $S_j'(x) = b_j + 2c_j(x - x_j) + 3d_j(x - x_j)^2$ y $S_j''(x) = 2c_j + 6d_j(x - x_j)$, entonces, por la continuidad de la primera derivada, tenemos

$$b_{j+1} = S_{j+1}'(x_{j+1}) = S_j'(x_{j+1}) = b_j + 2c_j h_j + 3d_j h_j^2, \quad j = 0, 1, \dots, n-1. \tag{4.2}$$

Análogamente por la continuidad de la segunda derivada

$$2c_{j+1} = S''_{j+1}(x_{j+1}) = S''_j(x_{j+1}) = 2c_j + 6d_jh_j \quad \Rightarrow \quad c_{j+1} = c_j + 3d_jh_j, \quad j = 0, 1, \ldots, n-1. \tag{4.3}$$

Despejando ahora de la igualdad anterior d_j y sustituyendo su valor en (4.1), después de reordenar esta ecuación, se sigue que

$$b_j = \frac{1}{h_j}(a_{j+1} - a_j) - \frac{h_j}{3}(c_{j+1} + 2c_j), \quad j = 0, 1, \ldots, n-1,$$

Análogamente, despejando d_j de la igualdad (4.3) y sustituyendo su valor en (4.2), obtenemos

$$b_{j+1} = b_j + h_j(c_{j+1} - 2c_j), \quad j = 0, 1, \ldots, n-1.$$

Finalmente, combinando las ecuaciones anteriores, tenemos

$$h_jc_j + 2(h_j + h_{j+1})c_j + h_{j+1}c_{j+2} = \frac{3}{h_{j+1}}(a_{j+2} - a_{j+1}) - \frac{3}{h_j}(a_{j+1} - a_j), \quad j = 0, 1, \ldots, n-2.$$

El cálculo de c_j se realiza resolviendo el sistema lineal $A\mathbf{x} = \mathbf{b}$, de orden $n+1$, donde $\mathbf{x} = (c_0, c_1, \ldots, c_n)^T$ es el vector de las incognitas, $\mathbf{b}$ el vector de los términos independientes y A la matriz de coeficientes (matriz tridiagonal estrictamente dominante).

Las expresiones de la matriz A y el vector $\mathbf{b}$ que se obtienen para el *spline* cúbico natural son ([3])

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 & \ldots & 0 \\ h_0 & 2(h_0+h_1) & h_1 & 0 & \ldots & 0 \\ 0 & h_1 & 2(h_1+h_2) & h_2 & \ldots & 0 \\ \vdots & & & \ddots & \ddots & \vdots \\ 0 & & \ddots & & & 0 \\ 0 & 0 & \ldots & h_{n-2} & 2(h_{n-2}+h_{n-1}) & h_{n-1} \\ 0 & 0 & \ldots & 0 & 0 & 1 \end{pmatrix},$$

$$\mathbf{b} = \begin{pmatrix} 0 \\ \frac{3}{h_1}(a_2 - a_1) - \frac{3}{h_0}(a_1 - a_0) \\ \vdots \\ \frac{3}{h_{n-1}}(a_n - a_{n-1}) - \frac{3}{h_{n-2}}(a_{n-1} - a_{n-2}) \\ 0 \end{pmatrix},$$

y para el *spline* cúbico sujeto

$$A = \begin{pmatrix} 2h_0 & h_0 & 0 & 0 & \ldots & 0 \\ h_0 & 2(h_0+h_1) & h_1 & 0 & \ldots & 0 \\ 0 & h_1 & 2(h_1+h_2) & h_2 & \ldots & 0 \\ \vdots & & & \ddots & \ddots & \vdots \\ 0 & & \ddots & & & 0 \\ 0 & 0 & \ldots & h_{n-2} & 2(h_{n-2}+h_{n-1}) & h_{n-1} \\ 0 & 0 & \ldots & 0 & h_{n-1} & 2h_{n-1} \end{pmatrix},$$

$$\mathbf{b} = \begin{pmatrix} \frac{3}{h_0}(a_1 - a_0) - 3f'(a) \\ \frac{3}{h_1}(a_2 - a_1) - \frac{3}{h_0}(a_1 - a_0) \\ \vdots \\ \frac{3}{h_{n-1}}(a_n - a_{n-1}) - \frac{3}{h_{n-2}}(a_{n-1} - a_{n-2}) \\ 3f'(b) - \frac{3}{h_{n-1}}(a_n - a_{n-1}) \end{pmatrix}.$$

Por las propiedades de la matriz A, los sistemas de ecuaciones lineales se pueden resolver, en ambos casos, mediante el método de eliminación de Gauss sin pivoteo o la factorización LU de Doolitte. Y para el caso de sistemas grandes (n grande), tanto el método de Jacobi como el de Gauss-Seidel convergen a la solución exacta (siendo este último más rápido).

Ejemplo. Construimos un *spline* cúbico sujeto que se ajuste a la tabla

x	0	1	2	3
f	0	1/2	2	3/2

(4.4)

y que verifique las condiciones $S'(0) = 1/5$ y $S'(3) = -1$. Calculamos primero las siguientes igualdades:

$$h_0 = h_1 = h_2 = 1,$$

$$a_0 = f(x_0) = 0, \quad a_1 = f(x_1) = \frac{1}{2}, \quad a_2 = f(x_2) = 2, \quad a_3 = f(x_3) = \frac{3}{2}.$$

Como $n = 3$, tenemos

$$\begin{pmatrix} 2 & 1 & 0 & 0 \\ 1 & 4 & 1 & 0 \\ 0 & 1 & 4 & 1 \\ 0 & 0 & 1 & 2 \end{pmatrix} \begin{pmatrix} c_0 \\ c_1 \\ c_2 \\ c_3 \end{pmatrix} = \begin{pmatrix} \frac{3}{2} - \frac{3}{5} \\ \frac{9}{2} - \frac{3}{2} \\ -\frac{3}{2} - \frac{9}{2} \\ -3 + \frac{3}{2} \end{pmatrix} = \begin{pmatrix} \frac{9}{10} \\ 3 \\ -6 \\ -\frac{3}{2} \end{pmatrix},$$

cuya solución es

$$c_0 = -\frac{9}{50}, \quad c_1 = \frac{63}{50}, \quad c_2 = -\frac{93}{50}, \quad c_3 = \frac{9}{50}.$$

Ahora

$$\begin{aligned} b_0 &= \frac{a_1 - a_0}{h_0} - \frac{(2c_0 + c_1)h_0}{3} = \frac{1}{5}, & d_0 &= \frac{c_1 - c_0}{3h_0} = \frac{12}{25}, \\ b_1 &= \frac{a_2 - a_1}{h_1} - \frac{(2c_1 + c_2)h_1}{3} = \frac{32}{25}, & d_1 &= \frac{c_2 - c_1}{3h_1} = -\frac{26}{25}, \\ b_2 &= \frac{a_3 - a_2}{h_2} - \frac{(2c_2 + c_3)h_2}{3} = \frac{17}{25}, & d_2 &= \frac{c_3 - c_2}{3h_2} = \frac{17}{25}. \end{aligned}$$

Luego

$$S(x) = \begin{cases} S_0(x) = a_0 + b_0(x - x_0) + c_0(x - x_0)^2 + d_0(x - x_0)^3 \\ \quad = \frac{1}{5}x - \frac{9}{50}x^2 + \frac{12}{25}x^3, & 0 \le x \le 1, \\ S_1(x) = a_1 + b_1(x - x_1) + c_1(x - x_1)^2 + d_1(x - x_1)^3 \\ \quad = \frac{1}{2} + \frac{32}{25}(x - 1) + \frac{63}{50}(x - 1)^2 - \frac{26}{25}(x - 1)^3, & 1 \le x \le 2, \\ S_2(x) = a_2 + b_2(x - x_2) + c_2(x - x_2)^2 + d_2(x - x_2)^3 \\ \quad = 2 + \frac{17}{25}(x - 2) - \frac{93}{50}(x - 2)^2 + \frac{17}{25}(x - 2)^3, & 2 \le x \le 3. \end{cases}$$

En la figura 4.5 podemos ver la gráfica del *spline* cúbico anterior $S(x)$. □

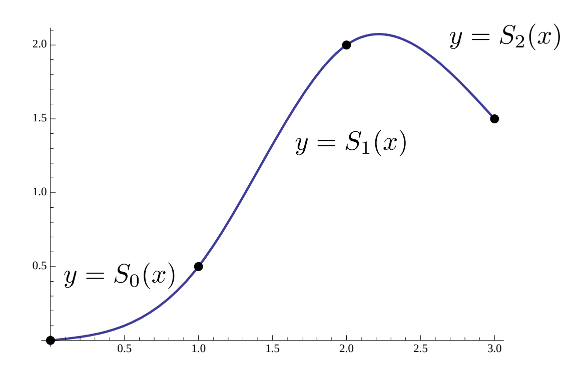


Figura 4.5: *Spline* cúbico sujeto $y = S(x)$ que interpola la tabla de datos 4.4.

4.3. Aproximación de mínimos cuadrados

Habitualmente los datos obtenidos a partir de experimentos conllevan errores sustanciales, de manera que la interpolación puede ser inadecuada y dar resultados poco satisfactorios cuando se utiliza para predecir valores intermedios. Una estrategia más adecuada en estos casos consiste en obtener una función de aproximación que se ajuste a la tendencia general de los datos, aunque no coincida en todos ellos. Diferentes funciones de aproximación se podrían entonces trazar, pero es conveniente dar algún criterio que sirva de base para el ajuste de los datos. Una opción consiste en encontrar una curva que minimice la discrepancia entre los datos y la curva. Un procedimiento que se puede utilizar para alcanzar dicho objetivo es el **método de mínimos cuadrados**.

4.3.1. Aproximación discreta de mínimos cuadrados

Sea un conjunto discreto de puntos $x_0, x_1, \ldots, x_n$ en el intervalo $[a,b]$ y sean $n+1$ valores reales y_i, de forma que se tienen los pares (x_i, y_i). Los puntos y_i han sido obtenidos mediante mediciones experimentales sobre cada punto x_i, o bien, porque cumplen $y_i = f(x_i)$ para una función $f(x)$ conocida al menos en los puntos x_i. Dado un conjunto de funciones conocidas $f_j(x)$, $j = 0, 1, \ldots, m$, definidas sobre $[a,b]$, queremos encontrar la función

$$u(x) = \sum_{j=0}^{m} a_j f_j(x), \quad a_j \in \mathbb{R},$$

que mejor se ajusta a los puntos dados, en el sentido de minimizar el error cuadrático

$$\sum_{i=0}^{n} (y_i - u(x_i))^2.$$

Nota. Las funciones $f_j(x)$, $j = 0, 1, \ldots, m$, definidas sobre $[a,b]$ se escogerán dependiendo de la forma de la nube de puntos (x_i, y_i), o bien, de las sospechas que tengamos respecto al comportamiento de $f(x)$. Por ejemplo, si observamos comportamientos periódicos, convendrá escoger como $f_j(x)$ funciones trigonométricas del tipo $\operatorname{sen}(jx)$ y $\cos(jx)$, $j = 0, 1, \ldots, m$; si observamos un comportamiento polinómico, escogeremos $f_j(x)$ como funciones polinómicas (por ejemplo, x^j, $j = 0, 1, \ldots, m$).

Como queremos que

$$\sum_{i=0}^{n} (y_i - u(x_i))^2 = \sum_{i=0}^{n} \left(y_i - \sum_{j=0}^{m} a_j f_j(x_i) \right)^2 = E(a_0, a_1, \ldots, a_m)$$

sea mínimo, tenemos que resolver, para cada $k = 0, 1, \ldots, m$, $\dfrac{\partial E}{\partial a_k} = 0$, i.e.:

$$-2 \sum_{i=0}^{n} \left(y_i - \sum_{j=0}^{m} a_j f_j(x_i) \right) f_k(x_i) = 0;$$

es decir, resolver

$$\sum_{i=0}^{n} \left(y_i - a_0 f_0(x_i) - \cdots - a_m f_m(x_i) \right) f_k(x_i) = 0, \quad k = 0, 1, \ldots, m.$$

Agrupando todos los a_j en cada ecuación, se tiene que se pueden escribir como:

$$a_0 \sum_{i=0}^{n} f_0(x_i) f_k(x_i) + a_1 \sum_{i=0}^{n} f_1(x_i) f_k(x_i) + \cdots + a_m \sum_{i=0}^{n} f_m(x_i) f_k(x_i) = \sum_{i=0}^{n} y_i f_k(x_i),$$

para $k = 0, 1, \ldots, m$. Si definimos la matriz $(n+1) \times (m+1)$:

$$M = \begin{pmatrix} f_0(x_0) & f_1(x_0) & \ldots & f_m(x_0) \\ f_0(x_1) & f_1(x_1) & \ldots & f_m(x_1) \\ \vdots & \vdots & \ddots & \vdots \\ f_0(x_n) & f_1(x_n) & \ldots & f_m(x_n) \end{pmatrix}$$

y los vectores $\mathbf{x} = (a_0, a_1, \ldots, a_m)^T$ e $\mathbf{y} = (y_0, y_1, \ldots, y_n)^T$, entonces las $m+1$ ecuaciones anteriores se pueden escribir como

$$M^T M\mathbf{x} = M^T\mathbf{y} \qquad \text{o} \qquad A\mathbf{x} = \mathbf{b},$$

donde $M^T M = A$ y $M^T\mathbf{y} = \mathbf{b}$, que es un sistema lineal de $m+1$ ecuaciones con $m+1$ incógnitas. La matriz A es simétrica por definición, y se tiene el siguiente resultado:

> Las funciones $f_j(x)$, $j = 0, 1, \ldots, m$, son linealmente independientes si y solo si la matriz A es definida positiva.

Por tanto, si tomamos $f_j(x)$ $(j = 0, 1, \ldots, m)$ linealmente independientes (en los casos trigonométrico y polinómico así son), se tiene que el sistema anterior tiene una única solución a_j, $j = 0, 1, \ldots, m$, que hace que la función $u(x) = \sum_{j=0}^{m} a_j f_j(x)$ tenga una magnitud de error de aproximación $\sum_{i=0}^{n}(y_i - u(x_i))^2$ mínima entre todas las funciones de este tipo.

La solución del sistema lineal puede entonces calcularse por cualquiera de los métodos vistos en el capítulo correspondiente, siendo especialmente adecuado el de Cholesky por ser A una matriz simétrica y definida positiva, o bien, si m es grande, el método iterativo de Gauss-Seidel por el mismo razonamiento. Además, véase [1], existen otros métodos, que no veremos, que se favorecen de que el sistema $A\mathbf{x} = \mathbf{b}$ se puede escribir de la forma $M^T M\mathbf{x} = M^T\mathbf{y}$.

Ejemplo. Vamos a aplicar la aproximación discreta de mínimos cuadrados para obtener la recta que se ajusta a la siguiente tabla de datos:

x	-1	1	3
y	6	1	11

(4.5)

Como $f_0(x) = 1$ y $f_1(x) = x$, tenemos que

$$M = \begin{pmatrix} f_0(-1) & f_1(-1) \\ f_0(1) & f_1(1) \\ f_0(3) & f_1(3) \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ 1 & 1 \\ 1 & 3 \end{pmatrix} \qquad \Rightarrow \qquad M^T = \begin{pmatrix} 1 & 1 & 1 \\ -1 & 1 & 3 \end{pmatrix},$$

de manera que

$$M^T M = \begin{pmatrix} 3 & 3 \\ 3 & 11 \end{pmatrix} = A, \qquad M^T y = \begin{pmatrix} 1 & 1 & 1 \\ -1 & 1 & 3 \end{pmatrix} \begin{pmatrix} 6 \\ 1 \\ 11 \end{pmatrix} = \begin{pmatrix} 18 \\ 28 \end{pmatrix} = \mathbf{b}$$

y el correspondiente sistema lineal $A\mathbf{x} = \mathbf{b}$ de dos ecuaciones con dos incógnitas tiene como única solución $\mathbf{x} = (a_0, a_1)^T = (4.75, 1.25)^T$. Por lo tanto, la recta de mínimos cuadrados que se obtiene es $u(x) = 4.75 + 1.25\,x$, que está representada en la figura 4.6. □

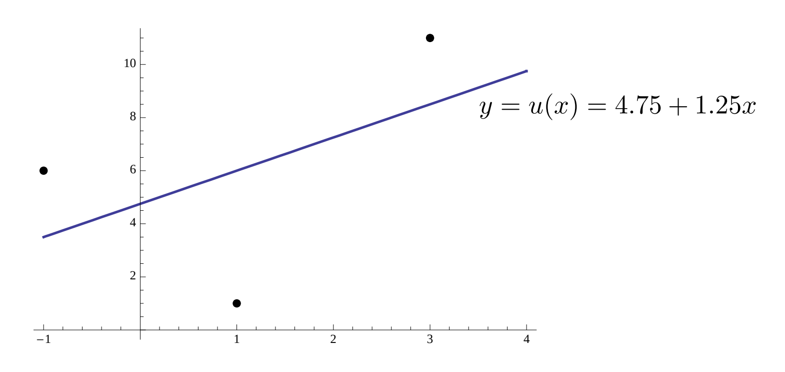


Figura 4.6: Aproximación discreta de mínimos cuadrados que ajusta la tabla de datos 4.5.

4.3.2. Aproximación continua de mínimos cuadrados

El método de mínimos cuadrados sirve también para aproximar no solo un conjunto discreto de puntos, sino también toda una función en un intervalo. Así, si queremos aproximar una función $f(x)$ en un intervalo

$[a, b]$ mediante el polinomio

$$p_m(x) = a_0 + a_1 x + \cdots + a_{m-1} x^{m-1} + a_m x^m,$$

tendremos que minimizar en este caso la función de error

$$E(a_0, a_1, \ldots, a_m) = \int_a^b (f(x) - p_m(x))^2 \, dx.$$

Derivando parcialmente respecto a los coeficientes del polinomio e igualando a cero se obtienen las ecuaciones

$$\frac{\partial E}{\partial a_k} = 2 \int_a^b (f(x) - p_m(x)) x^k \, dx = 0, \quad k = 0, 1, \ldots, m,$$

que nos llevan a un sitema lineal $A\mathbf{x} = \mathbf{b}$, donde $A = (a_{ij})$ es una matriz $(m+1) \times (m+1)$ con

$$a_{ij} = \int_a^b x^{i-1} x^{j-1} \, dx,$$

$\mathbf{x} = (a_0, a_1, \ldots, a_m)^T$ y $\mathbf{b}$ es el vector con coordenada i-ésima

$$b_i = \int_a^b x^{i-1} f(x) \, dx.$$

De nuevo A es simétrica y definida positiva y la solución del sistema lineal puede calcularse por cualquiera de los métodos vistos con anterioridad, siendo especialmente adecuado el de Cholesky por ser A simétrica y definida positiva, o bien, si m es grande, el método iterativo de Gauss-Seidel por el mismo motivo.

Ejemplo. Calculamos el polinomio de grado dos $p_2(x) = a_0 + a_1 x + a_2 x^2$ mejor aproximación de mínimos cuadrados de la función $f(x) = \operatorname{sen} \pi x$ en el intervalo $[0, 1]$. Los elementos de la matriz $A = (a_{ij})$ y del vector $\mathbf{b} = (b_i)$ serán:

$$\begin{array}{lll}
a_{11} = \int_0^1 1 \cdot 1 \, dx = 1, & a_{12} = \int_0^1 1 \cdot x \, dx = \frac{1}{2}, & a_{13} = \int_0^1 1 \cdot x^2 \, dx = \frac{1}{3}, \\
a_{21} = \int_0^1 x \cdot 1 \, dx = \frac{1}{2}, & a_{22} = \int_0^1 x \cdot x \, dx = \frac{1}{3}, & a_{23} = \int_0^1 x \cdot x^2 \, dx = \frac{1}{4}, \\
a_{31} = \int_0^1 x^2 \cdot 1 \, dx = \frac{1}{3}, & a_{32} = \int_0^1 x^2 \cdot x \, dx = \frac{1}{4}, & a_{33} = \int_0^1 x^2 \cdot x^2 \, dx = \frac{1}{5}, \\
b_1 = \int_0^1 1 \cdot \operatorname{sen} \pi x \, dx = \frac{2}{\pi}, & b_2 = \int_0^1 x \cdot \operatorname{sen} \pi x \, dx = \frac{1}{\pi}, & b_3 = \int_0^1 x^2 \cdot \operatorname{sen} \pi x \, dx = \frac{\pi^2-4}{\pi^3},
\end{array}$$

de forma que el sitema lineal $A\mathbf{x} = \mathbf{b}$ es

$$\begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{3} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} 2/\pi \\ 1/\pi \\ \frac{\pi^2-4}{\pi^3} \end{pmatrix},$$

cuya solución es

$$a_0 = \frac{12(\pi^2 - 10)}{\pi^3}, \qquad a_1 = \frac{60(12 - \pi^2)}{\pi^3}, \qquad a_2 = \frac{60(\pi^2 - 12)}{\pi^3}.$$

Por tanto, $p_2(x) = \dfrac{12(\pi^2 - 10)}{\pi^3} + \dfrac{60(12 - \pi^2)}{\pi^3} x + \dfrac{60(\pi^2 - 12)}{\pi^3} x^2$. Véase la figura 4.7. □

4.4. Sugerencias para seguir leyendo

Una de las referencias más completas con un tratamiento matemático avanzado de la interpolación y la aproximación es Davis (1975). También se pueden consultar [15], Ralston y Rabinowitz (2001) o Stoer y Bulirsch (2002). Una excelente descripción de la interpolación mediante funciones *splines* está dada en De Boer (2001). Para una introducción a la aproximación mediante funciones racionales, se pueden consultar Ralston y Rabinowitz (2001) o Stoer y Bulirsch (2002). Una gran variedad de problemas y diferentes enfoques de la teoría de aproximación pueden verse en Davis (1975) y Lorentz (2005). En [17] y Atkinson (1989) podemos ver una introducción a la interpolación trigonométrica. Una interesante discusión sobre la interpolación en dos dimensiones se encuentra en Lancaster y Salkaukas (1986).

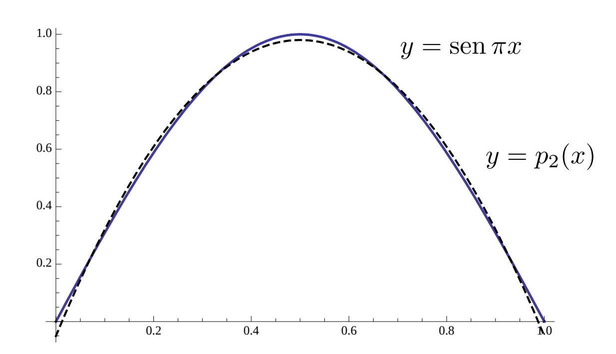


Figura 4.7: Función $y = \operatorname{sen} \pi x$ en el intervalo $[0, 1]$ (línea continua) y su polinomio $y = p_2(x)$ mejor aproximación de mínimos cuadrados (línea discontinua).

4.5. Ejercicios

1. *Interpolación de Taylor*. Dados los valores de la función f y sus derivadas sucesivas hasta orden n en un punto x_0 (es decir, $f^{(k)}(x_0)$ para $k = 0, 1, \ldots, n$), encuéntrese un polinomio p_n de grado $\leq\ n$ tal que $p_n^{(k)}(x_0) = f^{(k)}(x_0)$ para todo $k = 0, 1, \ldots, n$. Determínese que el problema de interpolación de Taylor tiene solución única. (Obviamente la función f ha de tener n derivadas en el punto x_0.)

2. Sea la siguiente tabla de la función $f(x) = \mathrm{e}^x$

x	0.0	0.2	0.4	0.6
$f(x)$	1.0000	1.2214	1.4918	1.8221

   a) Calcúlese $\sqrt[3]{\mathrm{e}}$ por interpolación cuadrática. Utilízese primero los puntos 0.0, 0.2 y 0.4 y posteriormente los puntos 0.2, 0.4 y 0.6. Compárense los resultados.

   b) Calcúlese $\sqrt[3]{\mathrm{e}}$ por interpolación cúbica y compárese este resultado con los anteriores y con el valor exacto, sabiendo que $\sqrt[3]{\mathrm{e}} = 1.395612425$.

3. El polinomio $p_3(x) = 2 - (x + 1) + x(x + 1) - 2x(x + 1)(x - 1)$ interpola los cuatro primeros datos de la tabla

x	-1	0	1	2	3
y	2	1	2	-7	10

   Añádase un término más a $p_3(x)$ de manera que el polinomio resultante interpole a todos los datos de la tabla. Calcúlese también el polinomio de Lagrange que interpola los datos de la tabla anterior.

4. Demuéstrese que los polinomios fundamentales de Lagrange, definidos por la expresión $\ell_k(x) = \prod_{j \neq k} \dfrac{x - x_j}{x_k - x_j}$ $(k = 0, \ldots, n)$, verifican $\sum_{k=0}^{n} \ell_k(x) = 1$.

5. Sea el polinomio de grado dos $f(x) = 3x^2 - x + 1$. Constrúyase el polinomio que interpola a f en los puntos $x_0 = 1$, $x_1 = 2$ y $x_2 = 3$. ¿Se vuelve a obtener el mismo polinomio f? ¿Por qué?

6. Sea el polinomio de grado tres $f(x) = 3x^3 - x + 1$. Utilizando diferencias divididas, constrúyase el polinomio que interpola a f en los puntos $x_0 = -3$, $x_1 = -1$ y $x_2 = 2$. Se añade a continuación un nuevo punto $x_3 = 1$. Sin hacer ningún cálculo adicional, razónese qué valor tomará la nueva diferencia dividida $f[x_0, x_1, x_2, x_3]$ y cuál será el coeficiente director (término de mayor grado) del nuevo polinomio de interpolación p_3 de f. Calcúlese p_3 según lo anterior. ¿Qué ocurre si consideramos $x_3 = 1.2$ en lugar del punto $x_3 = 1$?

7. *Importancia de la selección de puntos en la interpolación.* Como parece intuitivamente lógico, los puntos de interpolación deben estar centrados alrededor, y tan cerca como sea posible, de los valores deseados. Obsérvese esta realidad a la hora de aproximar el valor de $\ln 2 = 0.6931472$ después de interpolar la

función $f(x) = \ln x$ en los valores $x = 0, 4, 6, 5, 3, 1$ mediante polinomios de Newton de grados 1, 2 y 3. Estímese a continuación el error de cada aproximación. ¿Qué indican los resultados en relación con el grado del polinomio empleado para generar los datos? Justifíquese la respuesta.

8. Demuéstrese que los polinomios de interpolación de Lagrange y de Newton coinciden, es decir:

$$\sum_{i=0}^{n} f_i \ell_i(x) = f[x_0] + \sum_{i=1}^{n} \left(f[x_0, x_1, \ldots, x_i] \prod_{j=0}^{i-1} (x - x_j) \right).$$

9. Sea $f(x)$ la siguiente función definida en $[-1, 1]$

$$f(x) = \begin{cases} 0, & -1 \leq x \leq -0.25, \\ 1 - |4x|, & -0.25 \leq x \leq 0.25, \\ 0, & 0.25 \leq x \leq 1. \end{cases}$$

Encuéntrense polinomios de grados 2, 3, 4 y 5 que ajusten a f en puntos igualmente espaciados. Dibújense estas funciones y compruébese que el ajuste no es muy bueno. Finalmente, utilízense funciones *splines* cúbicas para ajustar la función anterior tomando como nodos cinco puntos equidistantes entre -1 y 1.

10. Sea la función $f(x) = \frac{1}{x}$ y los nodos $x_0 = 1$, $x_1 = 2$ y $x_2 = 3$.

    a) Constrúyase el polinomio de interpolación de Newton p_2 de f con los nodos dados.

    b) Determínense las constantes a_0, a_1, b_0, b_1 y b_3 para que la función

    $$S(x) = \begin{cases} a_0 + a_1(x-1) + \frac{13}{18}(x-1)^2 - \frac{2}{9}(x-1)^3, & 1 \leq x \leq 2, \\ b_0 + b_1(x-2) + \frac{1}{18}(x-2)^2 + b_3(x-2)^3, & 2 \leq x \leq 3, \end{cases}$$

    sea el *spline* cúbico con condiciones de contorno $S'(1) = -1$ y $S'(3) = -\frac{1}{9}$ que interpola a f en los nodos dados.

    c) Estímese el valor de $f(1.5)$ mediante $p_2(x)$ y $S(x)$. ¿Cuál de estas dos aproximaciones es mejor?

    d) Si añadimos un nodo más, $x_3 = 4$, determínese el polinomio de interpolación p_3 de f. ¿Se obtiene con $p_3(x)$ mejor aproximación del valor de $f(1.5)$ que las obtenidas en el apartado anterior?

11. Sea la tabla de datos

x	10	25	40	55
y	12	26	28	30

    a) Encuéntrese un *spline* lineal que interpole la tabla y evalúese en $x = 20$ y $x = 45$.

    b) Encuéntrese un *spline* cuadrático que interpole la tabla y evalúese en $x = 20$ y $x = 45$.

12. Demuéstrese que la solución obtenida $u(x)$ en la aproximación discreta de mínimos cuadrados coincide con el polinomio de interpolación $p_n(x)$ en los pares de puntos $(x_i, f(x_i))$ si $n = m$ y $f_j(x) = x^j$, $j = 0, 1, \ldots, m$.

13. Hállese el polinomio de segundo grado mejor aproximación de mínimos cuadrados para cada una de las siguientes funciones dadas en el intervalo indicado:

    a) $f(x) = \dfrac{1}{x}$ en $[1, 2]$, *b*) $f(x) = x^3 - 1$ en $[0, 2]$, *c*) $f(x) = \cos \pi x$ en $[0, 1]$, *d*) $f(x) = |x|$ en $[-1, 1]$.

14. La siguiente tabla de datos

x	-1	-0.75	-0.5	-0.25	0	0.25	0.5	0.75	1
y	-3.4738	0.4321	6.5455	-4.6710	0.0012	4.6797	-6.5510	-0.4375	3.4815

responde a una función del tipo $u(x) = a \operatorname{tg}(bx)$. Determínense a y b por el método de mínimos cuadrados. (Tómense 2.5 y 4.5 como valores iniciales aproximados de a y b respectivamente).

15. Encuéntrese la mejor aproximación de mínimos cuadrados en los puntos $(-\pi/2, 1)$, $(0, 0)$, $(\pi/2, 1/2)$ y $(\pi, 1)$ del tipo $u(x) = a + r\,\mathrm{sen}(x + \alpha)$, con a, r reales y $\alpha \in [0, 2\pi]$.

16. *Transformación de una ecuación no lineal en forma lineal.* A partir de los siguientes datos

x	1	2	5	9
y	4.1	3.4	1.8	0.8

parece claro que la mejor función que los aproxima es de tipo exponencial de la forma $u(x) = a\,\mathrm{e}^{bt}$. Determínense entonces los valores de a y b por el método de mínimos cuadrados. (*Indicación:* linealícese la función $u(x) = a\,\mathrm{e}^{bt}$ aplicandole el logaritmo natural para expresarla como un polinomio lineal y después transfórmese a la forma deseada.)

Capítulo 5

Derivación e integración numéricas

5.1. Introducción

En el capítulo anterior se ha puesto de manifiesto la utilidad del polinomio de interpolación de una función f cuando se quiere aproximar el valor de f en puntos en los que no se conoce dicho valor. En este capítulo lo utilizaremos para resolver dos problemas clásicos del cálculo numérico: la aproximación de derivadas e integrales definidas.

Es habitual que para una función genérica no siempre sea posible encontrar una primitiva de forma explícita; incluso si es conocida, puede ser difícil de utilizar. También es habitual que la función que se quiere integrar o derivar solo se conozca en un conjunto discreto de puntos (por ejemplo, cuando representa los resultados de mediciones experimentales), exactamente como sucede en el caso de la aproximación de funciones. En ambas situaciones es necesario considerar métodos numéricos para obtener una aproximación del valor que interesa, independientemente de lo difícil que sea la función a integrar o derivar. En estos casos la idea básica es la misma: aproximar la derivada de una función en un punto y la integral de una función en un intervalo, respectivamente, mediante la derivada de dicho punto de su polinomio de interpolación y la integral del polinomio de interpolación en dicho intervalo.

En primer lugar, daremos varias fórmulas para aproximar derivadas primeras y segundas mediante cocientes de diferencias. En segundo lugar, veremos varios métodos para aproximar una integral definida utilizando una suma ponderada de valores de la función en puntos específicos. Comenzaremos considerando fórmulas básicas que utilizan datos uniformemente espaciados, cuya exactitud será posteriormente mejorada a partir de la subdivisión del intervalo de integración y la aplicación de una de las técnicas básicas en cada subintervalo. Terminaremos presentando una técnica más potente, la *cuadratura gaussiana*, que utiliza una colección de puntos que son elegidos de forma que se obtenga el mejor resultado posible dentro de una cierta clase de funciones.

Destacamos que la aproximación de integrales (una tarea frecuentemente necesaria) puede normalmente llevarse a cabo sin mucho esfuerzo de manera muy precisa, mientras que la aproximación de derivadas (que es mucho menos necesaria) es un problema más difícil. Véase [7] para su análisis.

Necesitaremos posteriormente este tipo de métodos, cuando aproximemos la soluciones de ecuaciones diferenciales ordinarias y en derivadas parciales.

5.2. Derivación numérica

Las aproximaciones numéricas de las derivadas se usan principalmente de dos maneras. Una, cuando estamos interesados en calcular el valor de alguna derivada en algún punto prefijado, que a menudo se ha obtenido empíricamente. Dos, las fórmulas de derivación numérica se usan para obtener métodos numéricos en la resolución de ecuaciones diferenciales ordinarias y en derivadas parciales.

5.2.1. El problema de la derivación numérica

Nos planteamos el problema de aproximar la derivada de una función en un punto, bien porque solo conocemos una tabla de valores de la misma, o bien, porque la expresión de la función es excesivamente

complicada para intentar obtener una expresión explícita de su derivada.

Considérese una función $f : [a, b] \to \mathbb{R}$ continuamente diferenciable en $[a, b]$. En primer lugar, buscamos una aproximación de la derivada primera de f en un punto genérico c de (a, b). Lo lógico es aprovechar los valores conocidos de la función f para obtener un valor aproximado de $f'(c)$. Por ejemplo, como

$$f'(c) = \lim_{h\to 0} \frac{f(c+h) - f(c)}{h},$$

si se conoce el valor de f en c y $c + h$, cuando h es pequeño, la expresión

$$\frac{f(c+h) - f(c)}{h}$$

es una aproximación de $f'(c)$, véase la figura 5.1. Esta es una **fórmula de derivación numérica**, que además proporciona el valor exacto de $f'(c)$ cuando f es un polinomio de grado ≤ 1. Aunque esto puede parecer trivial, no resulta muy eficaz debido al error de redondeo. Sin embargo, es ciertamente el punto de partida.

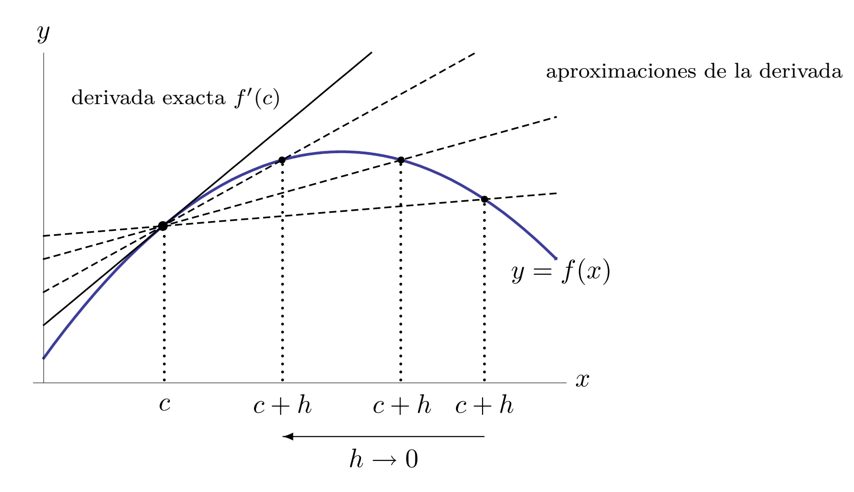


Figura 5.1: Derivada de una función en un punto c (recta continua) y aproximaciones de la derivada (rectas discontinuas).

5.2.2. Derivadas primeras

La derivación numérica de una función f diferenciable en $c \in \mathbb{R}$ consta de dos etapas:

- Construcción del polinomio de interpolación p_m de la función f en un conjunto de nodos $x_0, x_1, \ldots, x_n$ (que conviene tomar próximos a c).
- Derivación del polinomio p_n y evaluación en c, según la fórmula de derivación numérica: $f'(c) \simeq p_n'(c)$.

Las fórmulas así obtenidas reciben el nombre de **fórmulas de derivación interpolatorias**.

Para el caso del cálculo de la derivada primera de f en c, $f'(c)$, se pueden utilizar diferentes números de nodos. Con dos nodos, x_0 y x_1, se tiene que la fórmula de Newton del polinomo de interpolación es

$$p_1(x) = f(x_0) + f[x_0, x_1](x - x_0),$$

donde $f[x_0, x_1] = \dfrac{f(x_1) - f(x_0)}{x_1 - x_0}$. Por tanto, la correspondiente fórmula de derivación numérica es

$$f'(c) \simeq p_1'(c) = \frac{f(x_1) - f(x_0)}{x_1 - x_0}.$$

Es razonable que los nodos estén cerca de c. Si, por ejemplo, tomamos $x_0 = c$ y $x_1 = c + h$, con $h \neq 0$, la fórmula es

$$f'(c) \simeq \frac{f(c+h) - f(c)}{h}, \tag{5.1}$$

que se conoce como **diferencia finita progresiva** para la derivada primera. Se puede obtener otra aproximación de $f'(c)$ si tomamos $x_0 = c - h$ y $x_1 = c$, con $h \neq 0$, y la fórmula correspondiente

$$f'(c) \simeq \frac{f(c) - f(c-h)}{h},$$

se denomina entonces **diferencia finita regresiva** para la derivada primera. Estas dos últimas aproximaciones son *fórmulas de dos puntos*. Si $h \to 0$ en ambas fórmulas, la parte derecha tiende a $f'(c)$, puesto que es precisamente la definición de límite. Luego, para h pequeña, la aproximación anterior es buena (siempre y cuando no se produzcan errores de redondeo).

Si ahora tomamos $x_0 = c - h$ y $x_1 = c + h$, la fórmula resultante es

$$f'(c) \simeq \frac{f(c+h) - f(c-h)}{2h}. \tag{5.2}$$

NOTA. El *grado de exactitud* de una fórmula de derivación numérica es el mayor grado de los polinomios que son derivables exactamente por dicha fórmula. Se puede demostrar entonces que el grado de exactitud de la fórmula anterior de dos puntos es dos.

Con tres nodos x_0, x_1 y x_2, la fórmula de Newton del polinomo de interpolación es ahora

$$p_2(x) = f(x_0) + f[x_0, x_1](x - x_0) + f[x_0, x_1, x_2](x - x_0)(x - x_1),$$

donde $f[x_0, x_1, x_2] = \dfrac{f[x_1, x_2] - f[x_0, x_1]}{x_2 - x_0}$. Por lo tanto,

$$f'(c) \simeq p_2'(c) = \frac{f(x_1) - f(x_0)}{x_1 - x_0} + \frac{f[x_1, x_2] - f[x_0, x_1]}{x_2 - x_0}(c - x_1) + \frac{f[x_1, x_2] - f[x_0, x_1]}{x_2 - x_0}(c - x_0).$$

Teniendo en cuenta de nuevo que los nodos esten próximos a c, si $x_0 = c - h$, $x_1 = c$ y $x_2 = c + h$, se tiene

$$f'(c) \simeq \frac{f(c+h) - f(c-h)}{2h},$$

que coincide con (5.2) y se conoce como **diferencia finita centrada** para la derivada primera o **fórmula de tres puntos para el punto medio**.

Para $x_0 = c$, $x_1 = c + h$ y $x_2 = c + 2h$, se tiene

$$f'(c) \simeq \frac{-3f(c) + 4f(c+h) - f(c+2h)}{2h},$$

conocida como **diferencia finita progresiva de tres puntos**. Y si $x_0 = c - 2h$, $x_1 = c - h$ y $x_2 = c$, obtenemos

$$f'(c) \simeq \frac{3f(c) - 4f(c-h) + f(c-2h)}{2h},$$

que se conoce como **diferencia finita regresiva de tres puntos**.

Los métodos anteriores se llaman *fórmulas de tres puntos* (aunque el tercer punto $f(c)$ no aparezca en la fórmula para el punto medio). Análogamente, existen métodos conocidos como *fórmulas de cinco puntos* que involucran la evaluación de la función en dos nodos adicionales.

EJEMPLO. Dada $f(x) = \mathrm{e}^x$, vamos a aproximar $f'(1.5)$ mediante las fórmulas anteriores (5.1) y (5.2) con $h = 0.1$, y compararemos los resultados con el valor exacto $f'(x) = \mathrm{e}^{1.5}$. Si tomamos $x = 1.5$ y $h = 0.1$ en (5.1) y (5.2), tenemos respectivamente

$$f'(1.5) \simeq \frac{\mathrm{e}^{1.6} - \mathrm{e}^{1.5}}{0.1} \approx 4.713433540571 \qquad \text{y} \qquad f'(1.5) \simeq \frac{\mathrm{e}^{1.6} - \mathrm{e}^{1.4}}{0.2} \approx 4.489162287752.$$

Y los errores absolutos son entonces

$$\left|\mathrm{e}^{1.5} - 4.713433540571\right| = 0.231744 \qquad \text{y} \qquad \left|\mathrm{e}^{1.5} - 4.489162287752\right| = 0.007473. \quad \square$$

5.2.3. Derivadas segundas

Podemos generar métodos para aproximar las derivadas superiores de una función como hemos hecho para aproximar la derivada primera. Por ejemplo, si conocemos la función f en los nodos $x_0, x_1, \dots, x_m$, próximos a c, derivando k veces el polinomio de interpolación p_m y evaluando en c, tenemos $f^{(k)}(c) \simeq k!\, f[x_0, x_1, \dots, x_k]$.

A continuación vamos a ver algunas fórmulas para la derivada segunda. Debe haber tres o más nodos, ya que con dos el polinomio de interpolación p_1 es de primer grado, así que su derivada segunda es siempre nula para toda función f y todo punto c.

Con tres nodos x_0, x_1 y x_2, la fórmula de Newton del polinomo de interpolación es

$$p_2(x) = f(x_0) + f[x_0, x_1](x - x_0) + f[x_0, x_1, x_2](x - x_0)(x - x_1),$$

por lo que

$$f''(c) \simeq p_2''(c) = 2f[x_0, x_1, x_2] = 2\,\frac{f[x_1, x_2] - f[x_0, x_1]}{x_2 - x_0}.$$

Si $x_0 = c - h$, $x_1 = c$ y $x_2 = c + h$, se tiene

$$f''(c) \simeq \frac{f(c-h) - 2f(c) + f(c+h)}{h^2},$$

que se conoce como **diferencia finita centrada** para la derivada segunda o **fórmula de tres puntos para aproximar f'' en el punto medio**. Si tomamos $x_0 = c$, $x_1 = c + h$ y $x_2 = c + 2h$, entonces

$$f''(c) \simeq \frac{f(c) - 2f(c+h) + f(c+2h)}{h^2}.$$

Y si $x_0 = c - 2h$, $x_1 = c - h$ y $x_2 = c$,

$$f''(c) \simeq \frac{f(c-2h) - 2f(c-h) + f(c)}{h^2}.$$

Finalmente, mencionar que en los tres casos los métodos dan el valor exacto de la derivada si $h \to 0$.

Ejemplo. Sea $f(x) = \operatorname{sen} x$. Aproximemos el valor de $f''(0.5)$ mediante la diferencia finita centrada con $h = 0.1$, y comparémoslo con el valor real $\operatorname{sen}(0.5) = -0.47942554$. Así,

$$f''(0.5) \simeq \frac{f(0.4) - 2f(0.5) + f(0.6)}{(0.1)^2} \approx -0.4790027,$$

cuyo error absoluto es 0.000399. □

5.2.4. El error en la derivación numérica

Teniendo en cuenta la expresión del error de interpolación, es fácil dar expresiones del error de derivación numérica, pues basta con derivar la fórmula del error para el polinomio de interpolación. Sin embargo, en el caso de las fórmulas de derivación numérica también es fácil aprovechar el polinomio de Taylor de la función f para obtener una expresión del error de truncamiento. Por ejemplo, a partir de la fórmula

$$f(c+h) = f(c) + hf'(c) + \frac{h^2}{2!}f''(\xi), \quad \xi \in (c, c+h),$$

se deduce que

$$f'(c) - \frac{f(c+h) - f(c)}{h} = -\frac{h}{2!}f''(\xi), \quad \xi \in (c, c+h),$$

que da el error para la fórmula de dos puntos (5.1).

Análogamente, restando los desarrollos

$$f(c+h) = f(c) + hf'(c) + \frac{h^2}{2!}f''(c) + \frac{h^3}{3!}f'''(\xi_1), \quad \xi_1 \in (c, c+h),$$

$$f(c-h) = f(c) - hf'(c) + \frac{h^2}{2!}f''(c) - \frac{h^3}{3!}f'''(\xi_2), \quad \xi_2 \in (c-h, c),$$

se tiene

$$f'(c) - \frac{f(c+h) - f(c-h)}{2h} = -\frac{h^2}{12}(f'''(\xi_1) + f'''(\xi_2)) = -\frac{h^2}{6} f'''(\xi), \tag{5.3}$$

donde $\xi \in (c-h, c+h)$, siempre que f''' sea continua (para poder usar el teorema del valor intermedio y poder deducir entonces que existe el valor ξ), dando lugar así a la expresión del error asociado a la fórmula de derivación numérica (5.2). Si los valores de $f'''(\xi)$ no cambian muy rápidamente, entonces el error de truncamiento tiende a cero a la misma velocidad que h^2, lo que expresamos mediante la notación $\mathcal{O}(h^2)$.

De igual forma se obtienen expresiones para las restantes fórmulas de derivación numérica, aunque en algunos casos no es posible agrupar todos los restos de Taylor en un solo sumando. No obstante, en todas las fórmulas de derivación numérica, el error de truncamiento es proporcional a una potencia de h mayor o igual que uno.

Es especialmente importante prestar atención a los errores de redondeo cuando se aproximan derivadas. Cuando se aplica una fórmula de derivación numérica, el error de truncamiento disminuye si se reduce el tamaño de paso h, pero a costa de incrementar el error de redondeo. En la práctica, tomar h demasiado pequeño no suele reportar ventajas porque los errores de redondeo dominan los cálculos (véase la pág. 182 de [7] para un ejemplo). Todas las fórmulas de derivación numérica presentan problemas debidos al redondeo y, aunque los métodos de orden superior reduzcan las dificultades, es imposible evitar este problema enteramente.

Hay que tener en cuenta que, como método de aproximación, la derivación numérica es *inestable*, puesto que los valores pequeños de h que permitirían reducir el error de truncamiento aumentan los errores de redondeo. Esto debería evitarse si fuera posible, pero no lo es.

El valor óptimo de h es el que minimiza el error total $ET(h)$, es decir, la suma de los errores de truncamiento $E_t(h)$ y de redondeo $E_r(h)$. Así, para (5.3), si M es una cota de la tercera derivada de f, el error total $ET(h)$ verifica

$$|ET(h)| = |E_r(h) + E_t(h)| \leq |E_r(h)| + |E_t(h)| = \frac{\varepsilon}{h} + \frac{Mh^2}{6},$$

donde ε es la magnitud del error de redondeo máximo cometido al aproximar $f(c-h)$ y $f(c+h)$ con el ordenador. El mínimo del miembro de la derecha se alcanza cuando

$$-\frac{\varepsilon}{h^2} + \frac{Mh}{3} = 0; \quad \text{es decir,} \quad \text{para} \quad h = \sqrt[3]{\frac{3\varepsilon}{M}}.$$

Por tanto, este valor de h será bueno para aplicar dicha fórmula de derivación numérica, pero valores mucho menores que él pueden conducir a estimaciones pésimas de la derivada.

5.3. Integración numérica

En los cursos de cálculo se describen muchas técnicas para evaluar integrales exactamente, pero estas técnicas apenas pueden utilizarse para evaluar integrales que surgen en los problemas que se dan en la realidad. Las técnicas exactas no pueden resolver muchos problemas que aparecen en el mundo físico; para estos necesitamos métodos de aproximación de integrales. Estos métodos se llaman genéricamente **métodos de cuadratura** porque «cuadratura» es la palabra clásica para denominar el cálculo de áreas.

5.3.1. El problema de la cuadratura numérica

Estudiaremos a continuación métodos para el cálculo aproximado de integrales de la forma

$$\int_a^b f(x)\, dx,$$

donde f es una función integrable en el intervalo acotado $[a, b]$.

Ahora describimos tres situaciones en las que es necesario calcular aproximaciones a integrales definidas. La primera, el caso en el que la primitiva de la función f no se puede expresar en términos de funciones elementales. La segunda situación se debe al caso en que la primitiva se puede escribir, pero es tan complicada que se desea la aplicación de un método de cuadratura para su evaluación numérica. La tercera situación se da cuando el integrando se conoce solo puntualmente; por ejemplo, como resultado de una medición

experimental. El último caso también aparece cuando se aplican los métodos de cuadratura al tratamiento numérico de ecuaciones diferenciales o integrales.

Las mismas cuestiones que hemos visto para $f^{(k)}(c)$ pueden verse, sin apenas cambios, para aproximar $\int_a^b f(x)\,dx$, donde a y b son finitos y f una función definida e integrable en $[a,b]$. Para ello, aproximaremos f por un polinomio de interpolación p_m en un conjunto de nodos $x_0, x_1, \ldots, x_m$ y calcularemos exactamente $\int_a^b p_m(x)\,dx$, de manera que $\int_a^b f(x)\,dx \simeq \int_a^b p_m(x)\,dx$, obteniéndose así fórmulas de cuadratura numérica. Además, como los errores de interpolación tienen una fórmula explícita para las cotas del error, se pueden obtener también cotas del error para las fórmulas de cuadratura.

Empezamos introduciendo algunas fórmulas simples que son casos particulares de una familia mayor de fórmulas de cuadratura conocidas como *fórmulas de Newton-Cotes*. Para un conocimiento más completo de esta familia de fórmulas se puede consultar [1].

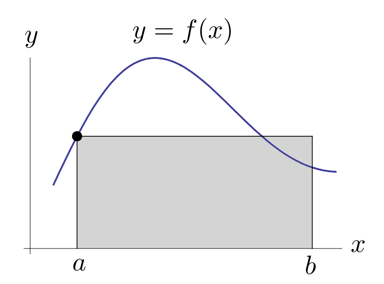
5.3.2. Reglas de cuadratura básicas

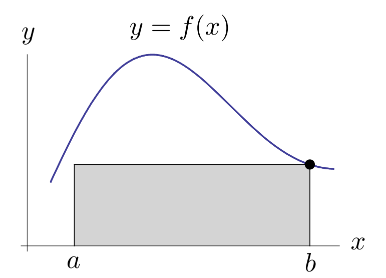
Comenzamos considerando un grupo de fórmulas de cuadratura numérica que están basadas en evaluaciones de funciones en nodos equiespaciados. Estas fórmulas se conocen como *fórmulas de Newton-Cotes*. Hay dos tipos básicos, que dependen de si los valores de la función en los extremos del intervalo de integración se utilizan o no. La regla del punto medio es el ejemplo más simple de fórmula de Newton-Cotes abierta, en la que los extremos de integración no se utilizan. Las reglas del trapecio y de Cavalieri-Simpson son ejemplos de fórmulas de Newton-Cotes cerradas, en las que los extremos de integración se utilizan.

El caso más simple que se puede dar es cuando solo se utiliza un nodo, x_0. En este caso, $p_0(x) = f(x_0)$, por lo que la integral en el intervalo $[a,b]$ se aproxima por

$$\int_a^b f(x)\,dx \simeq \int_a^b p_0(x)\,dx = \int_a^b f(x_0)\,dx = (b-a)f(x_0).$$

Gráficamente, si f es no negativa, lo que se hace es aproximar el área bajo la curva $y = f(x)$, comprendida entre $x = a$ y $x = b$, por el área del rectángulo de base $b - a$ y altura $f(x_0)$. Las elecciones más usuales son $x_0 = a$, $x_0 = b$ y $x_0 = \frac{a+b}{2}$. En el primer y segundo caso las fórmulas de cuadratura se llaman respectivamente **regla del rectángulo a izquierda** y **regla del rectángulo a derecha**. En el último casos la fórmula de cuadratura se llama **regla del punto medio**. En la figura 5.2 pueden verse las interpretaciones geométricas de estas tres reglas.





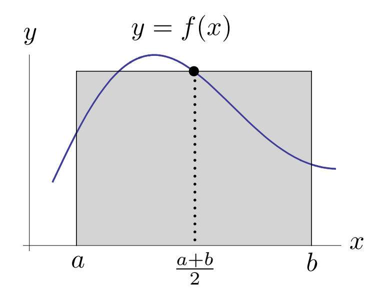


Figura 5.2: Regla del rectángulo a izquierda, regla del rectángulo a derecha y regla del punto medio.

El error para la regla del punto medio viene dado por la integral de la fórmula del error para el polinomio de interpolación p_0. Puede probarse, véase [7], que el error (de truncamiento local) de esta regla de cuadratura es: $E_{PM} = \frac{f''(\xi)}{24}(b-a)^3$, donde $\xi \in (a,b)$, siempre que $f \in C^2[a,b]$.

Si ahora utilizamos dos nodos, x_0 y x_1, el polinomio de interpolación es

$$p_1(x) = f(x_0) + f[x_0,x_1](x - x_0).$$

Por tanto la correspondiente fórmula de cuadratura será

$$\begin{aligned}
\int_a^b f(x)\,dx &\simeq \int_a^b p_1(x)\,dx \\
&= \int_a^b \left(f(x_0) + f[x_0,x_1](x-x_0)\right)\,dx \\
&= (b-a)f(x_0) + \frac{f(x_1) - f(x_0)}{x_1 - x_0}\left(\frac{(b-x_0)^2}{2} - \frac{(a-x_0)^2}{2}\right).
\end{aligned}$$

Si $x_0 = a$ y $x_1 = b$, se obtiene

$$\int_a^b f(x)\,dx \simeq \frac{b-a}{2}\,(f(a)+f(b)),$$

conocida como **regla del trapecio**. Geométricamente la regla del trapecio es equivalente a aproximar el área del trapecio bajo la recta que une $f(a)$ y $f(b)$, como puede verse en la figura 5.3.

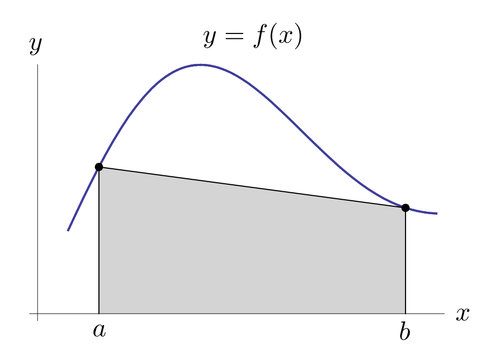


Figura 5.3: Regla del trapecio.

La expresión del error de la aproximación de la integral es (véase [7]): $E_T = -\frac{f''(\xi)}{12}(b-a)^3$, donde $\xi \in (a,b)$, siempre que $f \in C^2[a,b]$.

Cuando consideramos tres nodos x_0, x_1 y x_2, se tiene una de las fórmulas más importantes, la **regla de Cavalieri-Simpson**, que es la que corresponde a $x_0 = a$, $x_1 = \frac{a+b}{2}$ y $x_2 = b$. Si p_2 es el polinomo de interpolación en estos puntos, entonces

$$\int_a^b f(x)\,dx \simeq \int_a^b p_2(x)\,dx = \frac{b-a}{6}\left(f(a)+4f\left(\frac{a+b}{2}\right)+f(b)\right).$$

Geométricamente la regla de Cavalieri-Simpson es equivalente a aproximar el área de la figura bajo la parábola que pasa por $f(a)$, $f\left(\frac{a+b}{2}\right)$ y $f(b)$, como puede verse en la figura 5.4.

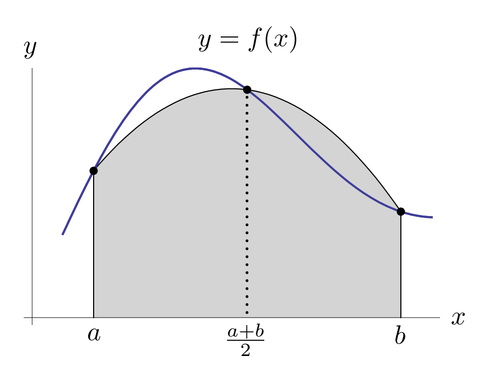


Figura 5.4: Regla de Cavalieri-Simpson.

El término de error para esta regla es: $E_S = -\frac{f^{(4)}(\xi)}{2880}(b-a)^5$, donde $\xi \in (a,b)$, siempre que $f \in C^4[a,b]$ (véase [25]).

Ejemplo. Para encontrar una aproximación de $\int_0^3 x^2\,\mathrm{e}^x\,dx$ mediante la regla de Cavalieri-Simpson, basta con

$$\int_0^3 x^2\,\mathrm{e}^x\,dx \simeq \frac{1}{2}\,(f(0)+4f(1.5)+f(3)) \approx 110.55252. \quad \square$$

Nota. El *grado de exactitud* de una fórmula de cuadratura es el mayor grado de los polinomios que son integrados exactamente por dicha fórmula. Se puede demostrar entonces que es uno para las reglas del punto medio y del trapecio, y tres para la regla de Cavalieri-Simpson.

COMENTARIO ADICIONAL. ▷ En todas las fórmulas anteriores, el error es proporcional a una potencia de la longitud del intervalo $(b-a)$. Por tanto, si el intervalo es pequeño, podemos esperar que el error sea pequeño, pero si es grande, no tenemos esa garantía. Obsérvese también que la elevada potencia de $(b-a)$ en la fórmula del error de la regla de Cavalieri-Simpson hace que ésta sea significativamente mejor que las reglas del punto medio y del trapecio, siempre que $b-a$ sea pequeño. Una ilustración de esto último puede verse en el ejemplo 1 de la pág. 123 de [7].

5.3.3. Reglas de cuadratura compuestas

En la sección anterior hemos tratado las nociones básicas que sustentan la integración numérica, pero las técnicas que hemos estudiado no son satisfactorias en muchos problemas. La hipótesis de que el intervalo sea pequeño para esperar un error pequeño podría ser muy poco razonable. No hay motivo, en general, para suponer que el intervalo $[a,b]$ sobre el que se integra es pequeño y, si no lo es, la elevada potencia de $(b-a)$ en la fórmula del error dominará, probablemente, los cálculos.

Resolveremos el problema de un intervalo de integración grande $[a,b]$ subdividiéndolo en una colección de intervalos que sean lo suficientemente pequeños para que podamos mantener bajo control el error en cada uno de ellos. Al sumar todos los resultados parciales se obtiene una fórmula de aproximación de la integral en $[a,b]$, dando lugar así a las llamadas **reglas de cuadratura compuestas**. El error en las fórmulas de cuadratura compuestas es entonces la suma de los errores de las fórmulas simples usadas en los subintervalos en los que se ha dividido $[a,b]$.

Un método de cuadratura compuesto consiste en dividir el intervalo $[a,b]$ en n subintervalos $[x_i,x_{i+1}]$ $(i=0,1,\dots,n-1)$ tal que

$$\int_a^b f(x)\,dx=\sum_{i=0}^{n-1}\int_{x_i}^{x_{i+1}} f(x)\,dx,$$

y aproximar ahora cada una de las integrales $\int_{x_i}^{x_{i+1}} f(x)\,dx$ mediante una fórmula de cuadratura básica.

Por ejemplo, si desomponemos el intervalo de integración en n subintervalos iguales mediante una partición uniforme del mismo, con nodos $x_i=a+ih$, $i=0,1,\dots,n$, $h=\frac{b-a}{n}$, y, en cada subintervalo, aproximamos por la fórmula del trapecio

$$\int_{x_i}^{x_{i+1}} f(x)\,dx\simeq\frac{h}{2}(f(x_i)+f(x_{i+1})),\quad i=0,1,\dots,n-1,$$

se obtiene la **regla del trapecio compuesta** para n subintervalos

$$\int_a^b f(x)\,dx\simeq\frac{h}{2}\sum_{i=0}^{n-1}(f(x_i)+f(x_{i+1}))=\frac{h}{2}\left(f(a)+f(b)+2\sum_{i=1}^{n-1}f(x_i)\right),$$

donde el error cometido viene dado por $E_{TC}=-\frac{h^2}{12}f''(\xi)(b-a)$, con $\xi\in(a,b)$, siempre que $f\in C^2[a,b]$.

El *orden de convergencia* de la regla del trapecio compuesta es entonces $\mathcal{O}(h^2)$, así que las aproximaciones convergen a $\int_a^b f(x)\,dx$ aproximadamente a la misma velocidad que $h^2\to 0$.

EJEMPLO. Utilizamos ahora la regla del trapecio compuesta con $n=2$ para aproximar la integral $\int_0^3 x^2\,\mathrm{e}^x\,dx$. Así, $h=3/2$ y

$$\int_0^3 x^2\,\mathrm{e}^x\,dx\simeq\frac{3}{4}\left(f(0)+f(3)\right)+\frac{3}{2}f(1.5)\approx 150.70307.$$

Notemos que el valor exacto de la integral anterior es $5\,\mathrm{e}^3-2\approx 98.42768$. Luego, se ha obtenido una aproximación peor que la dada anteriormente por la regla de Cavalieri-Simpson (véase el ejemplo anterior). Esto es debido a que no hemos tomado el número suficiente de trapecios para obtener una mejor aproximación. ¿Cuántos se deberían tomar entonces? A menudo, se calcula la integral varias veces incrementando el número de trapecios cada vez, y se para cuando la diferencia entre dos respuestas sucesivas es satisfactoriamente pequeña. □

COMENTARIOS ADICIONALES. ▷ El esquema de subdivisiones empleado anteriormente se puede aplicar con cualquiera de las reglas vistas en la sección anterior, dando lugar así a las correspondientes *reglas del punto medio compuesta* y *de Cavalieri-Simpson compuesta*. El orden de convergencia de estas dos

reglas es respectivamente $\mathcal{O}(h^2)$ y $\mathcal{O}(h^4)$. Esto prueba que el error de la regla de Cavalieri-Simpson tiende a cero más rápidamente que el error de las reglas del punto medio y del trapecio cuando $h \to 0$. Esta velocidad de convergencia es suficiente para la mayoría de los problemas habituales, suponiendo que el intervalo de integración se divide de manera que h sea pequeño.

▷ Una propiedad importante que comparten todas las reglas compuestas es su estabilidad con respecto a los errores de redondeo. Véase la pág. 134 de [7] para la regla de Cavalieri-Simpson compuesta.

▷ Si n es grande, los intervalos de integración $[x_i, x_{i+1}]$ son pequeños, con lo cual se espera que el error cometido, al aplicar una regla de cuadratura en cada uno de estos intervalos, sea pequeño, siendo el error total la suma de los errores cometidos en cada intervalo de integración. Se puede demostrar que si $n \to \infty$, el error tiende a 0.

5.3.4. Cuadratura gaussiana

Las reglas de cuadratura básicas se han construido integrando polinomios de interpolación. La fórmula del error del polinomio de interpolación de grado m contiene la derivada de orden $m+1$ de la función a aproximar. Puesto que la derivada $(m+1)$-ésima de todo polinomio de grado $\leq m$ es cero, al aplicar fórmulas de este tipo a dichos polinomios, obtenemos un resultado exacto.

Todas las reglas de cuadratura básicas utilizan valores de la función en puntos igualmente espaciados. Esto resulta conveniente a la hora de combinar las fórmulas para generar las reglas compuestas, pero esta restricción puede disminuir significativamente la exactitud de la aproximación. Véase la pág. 146 de [7] para un ejemplo con la regla del trapecio.

Además, las reglas de cuadratura básicas son ejemplos de una fórmula de cuadratura más general de la forma:

$$\int_a^b f(x)\,dx \simeq \sum_{i=0}^{m} w_i f(x_i).$$

Los números reales $\{w_i\}$ son los *pesos de cuadratura*, mientras que los puntos $\{x_i\}$ son los *nodos de cuadratura*. En la **cuadratura gaussiana** se consideran fórmulas de integración numérica como la anterior, pero utilizando nodos que no están igualmente espaciados en el intervalo; se eligen los nodos $\{x_i\}$ en el intervalo $[a,b]$ y los pesos $\{w_i\}$ de manera que se minimice el error que se espera obtener en la aproximación. Para minimizar este error esperable, vamos a suponer que la mejor elección de estos valores es la que produce resultados exactos para una clase más amplia de polinomios. Una elección adecuada de los $m+1$ nodos proporciona fórmulas de cuadratura numérica exactas para polinomios de grado $\leq 2m+1$, dando lugar así a las *fórmulas de cuadratura gaussiana.*

Los pesos $\{w_i\}$ de la fórmula anterior son arbitrarios y la única restricción sobre los nodos $\{x_i\}$ es que deben estar en el intervalo de integración $[a,b]$. Esto da $2(m+1)$ parámetros a elegir. Si ahora imponemos que sea exacta para los polinomios $1, x, x^2, \ldots, x^{2m+1}$ (que son polinomios de grado $\leq 2m+1$), obtenemos, mediante el método de los coeficientes indeterminados, un sistema no lineal de $2(m+1)$ ecuaciones con $2(m+1)$ incógnitas. Este sistema se puede resolver utilizando, por ejemplo, el método de Newton para sistemas de ecuaciones no lineales, o bien, convirtiéndolo, mediante ciertas transformaciones, en un sistema lineal y resolver éste con las técnicas correspondientes.

La utilización del método de los coeficientes indeterminados no resulta práctica para obtener fórmulas de cuadratura con grado de exactitud superior. Hay un método alternativo para obtener más fácilmente los nodos y los pesos de estas fórmulas que dan resultado exacto para polinomios de grado superior. Se consideran familias de polinomios especiales, llamados *polinomios ortogonales*, que tienen la propiedad de que una cierta integral definida del producto de dos polinomios ortogonales cualesquiera de la familia es cero.

La familia relevante para nuestro problema es la de los *polinomios de Legendre* en el intervalo $[-1,1]$. Estos polinomios pueden calcularse recursivamente mediante la siguiente relación de tres términos

$$\begin{aligned} &L_0(x) = 1, \quad L_1(x) = x, \\ &L_{k+1}(x) = \tfrac{2k+1}{k+1}\, x L_k(x) - \tfrac{k}{k+1}\, L_{k-1}(x), \quad k = 1, 2, \ldots \end{aligned}$$

Todo polinomio de grado $\leq m$, para cada $m \geq 0$, se puede obtener mediante una combinación lineal de los polinomio $L_0, L_1, \ldots, L_m$. El máximo grado de exactitud es $2m+1$ y se obtiene para la llamada **cuadratura**

de Gauss-Legendre, cuyos nodos y pesos están dados por

$$\begin{cases} x_i = \text{ceros de } L_{m+1}(x), \\ w_i = \dfrac{2}{(1-x_i^2)(L'_{m+1}(x_i))^2}, \end{cases} \quad i = 0, 1, \ldots, m.$$

Los pesos $\{w_i\}$ son todos positivos y los nodos $\{x_i\}$ son interiores al intervalo $(-1, 1)$. En la tabla 5.1 se recogen los nodos y pesos de las reglas de cuadratura de Gauss-Legendre con $m = 1, 2, 3, 4$. Para $m \geq 5$ se puede consultar [13]. Si $f \in C^{(2m+2)}([-1, 1])$, el correspondiente error es (véase [23]):

$$E_{GL} = \frac{2^{2m+3}((m+1)!)^4}{(2m+3)((2m+2)!)^3} f^{(2m+2)}(\xi), \quad \text{con} \quad \xi \in (-1, 1).$$

m	$\{x_i\}$	$\{w_i\}$
1	$\pm 1/\sqrt{3}$	1
2	$\pm\sqrt{15}/5$ 0	5/9 8/9
3	$\pm\frac{1}{35}\sqrt{525+70\sqrt{30}}$ $\pm\frac{1}{35}\sqrt{525-70\sqrt{30}}$	$\left(18-\sqrt{30}\right)/36$ $\left(18+\sqrt{30}\right)/36$
4	$\pm\frac{1}{21}\sqrt{245+14\sqrt{70}}$ $\pm\frac{1}{21}\sqrt{245-14\sqrt{70}}$ 0	$\left(322-13\sqrt{70}\right)/900$ $\left(322+13\sqrt{70}\right)/900$ 128/225

Tabla 5.1: Nodos y pesos para algunas fórmulas de cuadratura de Gauss-Legendre sobre el intervalo $(-1, 1)$. Los pesos correspondientes a pares simétricos de nodos se incluyen solo una vez.

Esto completa la solución del problema de aproximación de integrales definidas para funciones en el intervalo $[-1, 1]$. Ahora bien, esta solución es suficiente para cualquier intervalo cerrado porque la sencilla relación lineal

$$t = \frac{2x - a - b}{b - a} \qquad \Leftrightarrow \qquad x = \frac{b-a}{2}t + \frac{b+a}{2}$$

transforma la variable x del intervalo $[a, b]$ en la variable t del intervalo $[-1, 1]$. Entonces, podemos utilizar los polinomios de Legendre para aproximar

$$\int_a^b f(x)\,dx = \frac{b-a}{2}\int_{-1}^{1} f\left(\frac{b-a}{2}t + \frac{b+a}{2}\right)dt.$$

Ejemplo. Calculemos ahora un valor aproximado de la integral $\int_0^3 x^2\,\mathrm{e}^x\,dx$ mediante la cuadratura de Gauss-Legendre con $m = 2$. En primer lugar, transformamos la variable x del intervalo $[0, 3]$ en la variable t del intervalo $[-1, 1]$, de manera que

$$x = \frac{3}{2}(t+1) \qquad \text{y} \qquad \int_0^3 f(x)\,dx = \frac{3}{2}\int_{-1}^{1} f\left(\frac{3}{2}(t+1)\right)dt = \frac{3}{2}\int_{-1}^{1} g(t)\,dt,$$

donde $g(t) = f\left(\frac{3}{2}(t+1)\right) = \frac{9}{4}(t+1)^2\,\mathrm{e}^{3(t+1)/2}$. Por tanto, ya podemos utilizar los polinomios de Legendre para aproximar

$$\int_0^3 f(x)\,dx \simeq \frac{3}{2}\left(\frac{5}{9}\,g\left(-\frac{\sqrt{15}}{5}\right) + \frac{8}{9}\,g(0) + \frac{5}{9}\,g\left(\frac{\sqrt{15}}{5}\right)\right) \approx 98.15460.$$

Obsérvese que hemos obtenido así una mejor aproximación que las dadas anteriormente por las reglas de Cavalieri-Simpson y del trapecio compuesta. □

COMENTARIOS ADICIONALES. ▷ Para problemas pequeños, la regla de Cavalieri-Simpson compuesta puede ser aceptable para evitar la complejidad de los cálculos de las fórmulas de cuadratura gaussiana. Sin embargo, para problemas que requieran evaluaciones de funciones complicadas o realizar muchas evaluaciones de la integral, la eficiencia de la cuadratura gaussiana tiene una gran ventaja.

▷ Las fórmulas de cuadratura gaussiana son particularmente importantes para aproximar integrales múltiples, ya que el número de evaluaciones de la función crece como una potencia del número de integrales que se deben realizar. Véase [7].

▷ La cuadratura gaussiana no es apropiada cuando no se conoce la función porque requiere de evaluaciones de la función en puntos irregularmente espaciados dentro del intervalo de integración. Por ejemplo, si el problema tiene los datos tabulados, será necesario primero interpolar para los nodos considerados.

5.4. Sugerencias para seguir leyendo

Un texto muy interesante sobre cuadratura numérica es Davis y Rabinowitz (2007). Para las fórmulas básicas se pueden consultar [15] Atkinson (1989) o Ralston y Rabinowitz (2001).

5.5. Ejercicios

1. Sean $f(x) = x\,\mathrm{e}^x$ y

x_i	1.8	1.9	2.0	2.1	2.2
$f(x_i)$	10.88936	12.70319	14.77811	17.14896	19.85503

   a) Calcúlese $f'(x)$, evalúese $f'(2.0)$ y aproxímese $f'(2.0)$ mediante fórmulas de 2 y 3 puntos. ¿Qué fórmula obtiene mejor aproximación? ¿Por qué?

   b) Calcúlese $f''(x)$, evalúese $f''(2.0)$ y aproxímese $f''(2.0)$ mediante fórmulas de 3 puntos. ¿Qué fórmula obtiene mejor aproximación? ¿Por qué?

2. *a*) Demuéstrese que la diferencia finita centrada para la primera derivada es exacta para polinomios de grado ≤ 2, mientras que las diferencias finitas progresiva y regresiva de tres puntos son exactas para polinomios de grado ≤ 3.

   b) Demuéstrese que la diferencia finita centrada para la segunda derivada es exacta para polinomios de grado ≤ 3, mientras que las otras dos diferencias finitas presentadas en el texto son exactas para polinomios de grado ≤ 2.

3. Calcúlense los valores de a, b, c para que la fórmula de derivación numérica

$$f'(x) \simeq \frac{1}{h}(af(x+h) + bf(x) - cf(x-h))$$

   sea exacta para polinomios del mayor grado posible.

4. *Fórmula de cinco puntos para el punto medio.* Demuéstrese, a partir de los cinco nodos $c-2h$, $c-h$, c, $c+h$ y $c+2h$, que se puede obtener la siguiente fórmula de derivación numérica para la derivada primera

$$f'(x) \simeq \frac{f(c-2h) - 8f(c-h) + 8f(c+h) - f(c+2h)}{12h},$$

   y que es exacta para polinomios de grado ≤ 4.

5. Obténgase una fórmula de diferencias finitas para aproximar $f''(x)$ en los puntos $c-h$, c y $c+2h$. ¿Cuál es su grado de exactitud?

6. *Derivación parcial numérica.* Teniendo en cuenta que la derivada parcial $\frac{\partial f}{\partial x}(x,y)$ de $f(x,y)$ con respecto a x se calcula derivando con respecto a x y manteniendo fija y, y que la derivada parcial $\frac{\partial f}{\partial y}(x,y)$ de $f(x,y)$ con respecto a y se hace derivando con respecto a y y manteniendo fija x, todas las fórmulas de diferencias finitas vistas a lo largo del capítulo se pueden adaptar para aproximar derivadas parciales.

Adáptase entonces la fórmula (5.2) para calcular las derivadas parciales $\frac{\partial f}{\partial x}(x,y)$ y $\frac{\partial f}{\partial y}(x,y)$ y aplíquense las nuevas fórmulas a $f(x,y) = \frac{xy}{x+y}$ para aproximar $\frac{\partial f}{\partial x}(4,5)$ y $\frac{\partial f}{\partial x}(4,5)$ con $h = 0.1$ y $h = 0.01$. Compárense los resultados obtenidos con los valores exactos.

7. Obténgase la regla de Cavalieri-Simpson.

8. Demuéstrese que las reglas del punto medio y del trapecio son exactas para polinomios de grado ≤ 1 y que la regla de Cavalieri-Simpson lo es para polinomios de grado ≤ 3.

9. Determínese si la fórmula de Cavalieri-Simpson es exacta para las funciones $p(x) + \alpha \operatorname{sen} x$, donde $p(x)$ es un polinomio de grado ≤ 3 y $\alpha \in \mathbb{R}$, en el intervalo $[-\pi, \pi]$.

10. La solución aproximada de la integral:
$$\int_0^1 \mathrm{e}^{x^2}\, dx$$
es 1.46265. Calcúlese una aproximación de la integral dividiendo el intervalo de integración en cuatro subintervalos iguales y utilizando las reglas del punto medio, del trapecio y de Cavalieri-Simpson. Compárense los resultados.

11. Determínense los valores w, x_0 y x_1 para que la fórmula de cuadratura numérica
$$\int_0^1 f(x)\, dx \simeq w(f(x_0) + f(x_1))$$
sea exacta para polinomios del mayor grado posible. Utilícese la fórmula anterior para aproximar el valor de las siguientes dos integrales
$$\int_0^1 e^{x^2} dx \qquad \text{y} \qquad \int_{-1}^1 e^{x^2} dx.$$

12. Calcúlense los pesos w_i $(i = 0, 1, 2, 3)$ de manera que la fórmula de cuadratura numérica
$$\int_{-\pi}^{\pi} f(x) \cos x dx \simeq w_0 f\left(-\frac{3\pi}{4}\right) + w_1 f\left(-\frac{\pi}{4}\right) + w_2 f\left(\frac{\pi}{4}\right) + w_3 f\left(\frac{3\pi}{4}\right)$$
sea exacta para polinomios del mayor grado posible.

13. Determínense los valores w_0, w_1, x_0 y x_1 para que la fórmula de cuadratura numérica
$$\int_{-1}^1 f(x)(1 + x^2)\, dx \simeq w_0 f(x_0) + w_1 f(x_1)$$
sea exacta para polinomios del mayor grado posible. Aplíquese la fórmula a $f(x) = 2x^2$.

14. La solución aproximada de la integral:
$$\int_{-1}^1 \frac{\cos x}{\sqrt{1 - x^2}}\, dx$$
es 2.40394. Calcúlese una aproximación mediante la fórmula de Gauss-Legendre con $m = 2$ y $m = 4$. Compárense los resultados.

15. Calcúlese una aproximación de la integral del ejercicio 10, transformando previamente el intervalo de integración, mediante la fórmula de Gauss-Legendre con $m = 2$ y $m = 4$. Compárense los resultados.

16. Obténgase la fórmula de cuadratura de Gauss-Legendre de dos puntos $(m = 1)$ utilizando el método de los coeficientes indeterminados y la exactitud de la fórmula.

Capítulo 6

Resolución numérica de problemas de valor inicial

6.1. Introducción

Una ecuación diferencial es una ecuación en la que aparece una o más derivadas de una función desconocida. Si todas las derivadas se toman con respecto a una sola variable independiente se llama *ecuación diferencial ordinaria* (abreviadamente, EDO), mientras que hablaremos de una *ecuación en derivadas parciales* (abreviadamente, EDP) cuando aparecen derivadas parciales en la ecuación. Una ecuación diferencial (EDO o EDP) tiene *orden* q si q es el máximo orden de derivación que aparece en la ecuación. En este capítulo nos centraremos únicamente en las EDO de primer orden.

Las EDO modelizan una gran cantidad de fenómenos en diversos campos. En muchas situaciones reales un problema está modelizado por una EDO que requiere del cálculo de la solución de un *problema de valor inicial* (abreviadamente, PVI); esto es, la solución de una EDO que verifica una condición inicial dada. Frecuentemente el PVI es demasiado complicado como para que se pueda resolver exactamente, de manera que se emplea entonces una de dos posibles técnicas para aproximar su solución. La primera técnica consiste en simplificar la EDO, obteniendo otra que pueda resolverse exactamente, y usar después la solución de la ecuación simplificada como aproximación de la solución de la ecuación original. La otra técnica, que es la que trataremos aquí, consiste en construir métodos que aproximen directamente la solución del problema original.

Los métodos que consideraremos no proporcionan una aproximación continua a la solución del PVI, sino aproximaciones del valor de la solución en un conjunto de puntos que habitualmente están igualmente espaciados, de manera que será necesario utilizar algún tipo de interpolación cuando se necesiten valores intermedios. Las fórmulas de diferenciación numérica presentadas en el capítulo anterior juegan un papel fundamental en la construcción de estos métodos.

Los métodos de resolución que presentamos son incrementales, de manera que la solución se determina en un número de pasos. Empiezan en el punto en el que está dada la condición inicial. Después, utilizando la aproximación de la solución conocida en el primer punto, se determina una aproximación de la solución en un segundo punto más cercano. A continuación, se determina la aproximacíón de la solución en un tercer punto, y así sucesivamente. Existen *métodos de un paso* y *de varios pasos* (*multipaso*). En los métodos de un paso se aproxima la solución a partir de la aproximación en el paso anterior, mientras que en los métodos multipaso la solución se aproxima a partir de aproximaciones conocidas en varios pasos anteriores. El interés de los métodos multipaso es que el conocimiento del valor de la función en varios puntos anteriores puede dar una mejor estimación de la tendencia de la solución. Para aproximar la solución en cada paso también se pueden utilizar dos tipos de métodos: *métodos explícitos* y *métodos implícitos*. Los segundos suelen proporcionar mejor exactitud que los primeros, pero requieren de un mayor esfuerzo computacional en cada paso.

Limitaremos nuestro estudio a EDO de primer orden, dado que una ecuación de orden $q > 1$ siempre se puede reducir a un sistema de q ecuaciones de primer orden. El caso de sistemas de primer orden se tratará a continuación.

6.2. El problema de Cauchy

Una EDO admite, en general, un número infinito de soluciones. Para fijar una de ellas, debemos imponer una condición adicional que dée el valor tomado por esta solución en un punto dado del intervalo de integración. Por ejemplo, la ecuación $y'(t) = C_1(y - C_2)$ admite la familia de soluciones $y(t) = C_2 + K\,\mathrm{e}^{C_1 t}$, donde K es una constante arbitraria. Si imponemos la condición $y(0) = 1$, escogemos entonces la única solución correspondiente al valor $K = 1 - C_2$.

El **problema de Cauchy**, también llamado **problema de valor inicial**, consiste en encontrar la función solución de una EDO que satisfaga una condición inicial dada, y toma la siguiente forma:

> Encuéntrese una función real $y : [a, b] \to \mathbb{R}$ tal que
>
> $$y'(t) = f(t, y(t)), \text{ para todo } t \in [a, b], \text{ con } y(a) = y_0, \tag{6.1}$$
>
> donde $[a, b]$ es un intervalo de $\mathbb{R}$, $f : [a, b] \times \mathbb{R} \to \mathbb{R}$ es una función dada e y_0 es un valor dado que se llama *dato inicial*.

Antes de describir los métodos para aproximar la solución de nuestro problema básico, consideraremos algunas condiciones que garanticen que exista una solución. De hecho, puesto que no resolveremos el problema dado, sino solo una aproximación de él, necesitamos saber cuando un problema cercano al dado posee soluciones que se aproximan con precisión a la solución del problema original. Cuando un PVI posee esta propiedad se dice que está *bien planteado*, siendo éstos son los problemas para los que son adecuados los métodos numéricos. La siguiente **condición de buen planteamiento** establece que la clase de los problemas bien planteados es bastante amplia:

> Supongamos que f y f_y, su derivada parcial con respecto a y, son continuas para todo $t \in [a, b]$ y para todo y. Entonces, el problema de Cauchy (6.1) tiene *solución única* $y = y(t)$ para $t \in [a, b]$ y es un problema bien planteado.

La estrategia común de los métodos numéricos capaces de aproximar la solución de cada EDO para la que existe solución consiste en subdividir el intervalo de integración $[a, b]$ en N intervalos de longitud $h = \frac{b-a}{N}$; h se llama *paso de discretización* o *tamaño de paso*. Entonces, en cada *nodo* $t_i = a + ih$ $(i = 0, 1, \ldots, N)$, buscamos el valor desconocido u_i que aproxima a $y_i = y(t_i)$. El conjunto de valores $\{u_0 = y_0, u_1, \ldots, u_N\}$ es la solución numérica buscada.

6.3. Métodos de Taylor

Muchos de los métodos numéricos vistos hasta ahora se derivan en el fondo del Teorema de Taylor. La aproximación de la solución de un PVI no es una excepción. En este caso, la función que expresaremos en términos de su polinomio de Taylor es la solución (desconocida) del problema $y(t)$. Su forma más elemental (polinomio de Taylor con $n = 1$) conduce al **método de Euler** *explícito*, que genera una solución numérica como la que sigue:

$$u_0 = y_0, \quad u_{i+1} = u_i + hf(t_i, u_i), \quad i = 0, 1, \ldots, N - 1,$$

aproximándose así la EDO por una *ecuación en diferencias*. Algebraicamente este método se puede obtener considerando la EDO (6.1) en cada nodo $i = 1, 2, \ldots, N$, reemplazando la derivada exacta $y'(t_i)$ por la diferencia finita progresiva $\frac{1}{h}(y(t_{i+1}) - y(t_i))$ y construyéndose la aproximación u_i de $y(t_i)$ para cada $i = 1, 2, \ldots, N$.

Una interpretación geométrica de este método está dada en la figura 6.1 y se puede explicar como sigue. Supongamos que hemos encontrado u_i en $t = t_i$. La ecuación de la recta tangente a la gráfica de $y(t)$ en $t = t_i$ es $y - u_i = m(t - t_i)$, donde $m = y'(t_i) = f(t_i, u_i)$. Para $t = t_{i+1}$ e $y = u_{i+1}$, tenemos $u_{i+1} - u_i = m(t_{i+1} - t_i)$. Entonces, $u_{i+1} = u_i + hf(t_i, u_i)$. Esto muestra que la siguiente aproximación u_{i+1} se obtiene en el punto donde la tangente a la gráfica de $y(t)$ en $t = t_i$ interseca con la recta vertical $t = t_{i+1}$.

Se define el *error (de truncamiento) local* como el error que se comete en un paso cuando se supone que todos los resultados previos son exactos. El error verdadero, o acumulado, del método se denomina *error (de truncamiento) global* o *total*.

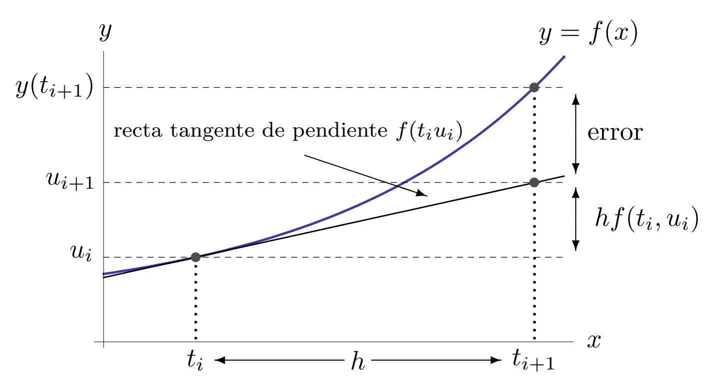


Figura 6.1: Interpretación geométrica del método de Euler.

Suponiendo que exista la derivada segunda de y y que es continua, obtenemos por el Teorema de Taylor (véase el error en la derivación numérica) que

$$y(t_{i+1}) - (y(t_i) + hf(t_i, y(t_i))) = \frac{h^2}{2} y''(\xi_i), \quad \text{para un } \xi_i \in (t_i, t_{i+1}) \text{ adecuado,}$$

siendo el error local en cada paso proporcional a h^2, así que es de orden $\mathcal{O}(h^2)$. Sin embargo, el error global acumula estos errores locales, así que generalmente crece mucho más rápidamente. Damos a continuación una **cota del error global para el método de Euler**:

Sea $y(t)$ la única solución del problema de Cauchy (6.1) y sean $u_0, u_1, \ldots, u_N$ las aproximaciones generadas por el método de Euler para algún número natural N. Supongamos que f es continua para todo $t \in [a, b]$ y todo $y \in (-\infty, +\infty)$, y que existen constantes L y M tales que

$$\left| \frac{\partial f}{\partial y}(t, y(t)) \right| \leq L, \qquad |y''(t)| \leq M.$$

Entonces, para cada $i = 0, 1, \ldots, N$, se tiene

$$|y(t_i) - u_i| \leq \frac{hM}{2L} \left(\mathrm{e}^{L(t_i - a)} - 1 \right).$$

Comentarios adicionales. ▷ Un aspecto importante que debemos resaltar es que, aunque el error local del método de Euler es de orden $\mathcal{O}(h^2)$, su error global es de orden $\mathcal{O}(h)$. Diremos entonces que el método de Euler es un método de primer orden.

▷ La reducción en una potencia del error local al error global es típica de las técnicas numéricas para PVI. De todas formas, aunque hay una reducción de orden del error local al global, la fórmula muestra que se puede reducir el error disminuyendo el paso, de manera que el error tiende a cero con h.

Ejemplo. Aplicamos el método de Euler con $N = 10$ para resolver el problema de valor inicial dado por:

$$y'(t) = 2t - y(t), \quad y(0) = -1,$$

y obtener el valor de y en $t = 1$. Comparamos también los resultados obtenidos con los valores de la solución exacta $y(t) = \mathrm{e}^{-t} + 2t - 2$.

Como $h = \frac{b-a}{N} = \frac{1}{10} = 0.1$ y $f(t, y) = 2t - y$, se sigue que

$$y(0.1) \simeq u_1 = u_0 + hf(t_0, u_0) = -1 + (0.1)f(0, -1) = -1 + 0.1 = -0.9,$$

y el error es $|y(0.1) - u_1| = 0.00484$. Para calcular $y(0.2)$, repetimos el proceso, pero empezando ahora en el punto $(0.1, -0.9)$:

$$y(0.2) \simeq u_2 = u_1 + hf(t_1, u_1) = -0.9 + (0.1)f(0.1, -0.9) = -0.9 + (0.1)(0.2 + 0.9) = -0.79;$$

el error absoluto es $|y(0.2) - u_2| = 0.00973$. Continuando el proceso de forma análoga, obtenemos, véase [16], que el valor de y en $t = 1$ es $u_{10} = 0.348678$ y el error cometido 0.0192. Finalizamos la aplicación del método de Euler mostrando las gráficas de las soluciones exacta y aproximada del PVI anterior en la figura 6.2. □

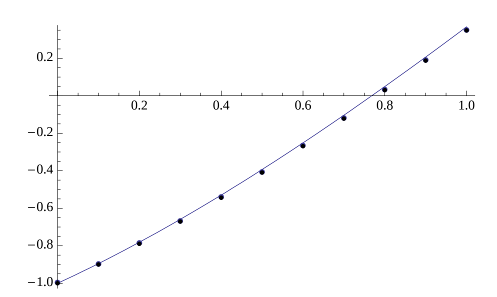


Figura 6.2: Soluciones exacta $y(t) = \mathrm{e}^{-t} + 2t - 2$ (línea continua) y aproximada mediante el método de Euler con $N = 10$ (puntos).

Puesto que se ha construido el método de Euler a partir del Teorema de Taylor, el primer intento para hallar métodos que mejoren la precisión del método de Euler consistirá en extender esta técnica de construcción. Así, véase [7], si expresamos la solución $y(t)$ del problema de Cauchy (6.1) en términos de su n-ésimo polinomio de Taylor alrededor de t_i, se deriva sucesivamente la solución $y(t)$ y se sustituyen estos resultados en el desarrollo de Taylor se obtiene, eliminando el término de error, la correspondiente ecuación en diferencias del **método de Taylor orden** n:

$$u_0 = y_0, \quad u_{i+1} = u_i + hT_i^{(n)}, \quad i = 0, 1, \dots, N-1, \tag{6.2}$$

donde

$$T_i^{(n)} = T^{(n)}(t_i, u_i) = f(t_i, u_i) + \frac{h}{2} f'(t_i, u_i) + \cdots + \frac{h^{n-1}}{n!} f^{(n-1)}(t_i, u_i).$$

El error local es $\frac{h^{n+1}}{(n+1)!} y^{(n+1)}(\xi_i)$, para algún $\xi_i \in (t_i, t_{i+1})$.

Las estimaciones del error global para los métodos de Taylor son parecidas a las del método de Euler. Si se dan suficientes condiciones de derivabilidad, entonces el método de Taylor de orden n tendrá un error local de $\mathcal{O}(h^{n+1})$ y un error global de orden $\mathcal{O}(h^n)$.

NOTA. La fórmula de $T^{(n)}$ se expresa fácilmente, pero es difícil de usar porque requiere conocer las derivadas de $f(t, y(t))$ con respecto a t. Como f es una función de las dos variables t e y, la regla de la cadena dice que la derivada total de f con respecto a t, que hemos denotado por $f'(t, y(t))$, está dada por

$$f'(t, y(t)) = \frac{\partial f}{\partial t}(t, y(t)) \cdot \frac{dt}{dt} + \frac{\partial f}{\partial y}(t, y(t)) \; \frac{dy(t)}{dt} = \frac{\partial f}{\partial t}(t, y(t)) + \frac{\partial f}{\partial y}(t, y(t)) \; y'(t),$$

o bien, puesto que $y'(t) = f(t, y(t))$, por

$$f'(t, y(t)) = \frac{\partial f}{\partial t}(t, y(t)) + f(t, y(t)) \; \frac{\partial f}{\partial y}(t, y(t)) = f_t(t, y(t)) + f(t, y(t)) f_y(t, y(t)).$$

Las derivadas de órdenes superiores se obtienen de forma parecida, pero se van haciendo cada vez más complicadas. Por ejemplo, $f''(t, y(t))$ involucra todas las derivadas parciales, tanto con respecto a t como a y, de todos los términos del miembro derecho de la última igualdad.

6.4. Métodos de Runge-Kutta

Acabamos de ver cómo se pueden generar métodos de Taylor de orden superior. Sin embargo, la aplicación de estos métodos de orden superior a problemas concretos se complica por la necesidad de calcular y evaluar las derivadas de orden superior con respecto a t del segundo miembro de la EDO (6.1).

Los métodos anteriores son ejemplos elementales de métodos de un paso. Esquemas más sofisticados, que permiten alcanzar un orden de precisión superior, son los *métodos de Runge-Kutta* y los *métodos multipaso*. En esta sección consideraremos los **métodos de Runge-Kutta** (abreviadamente, RK), que son métodos de un paso; sin embargo, requieren de varias evaluaciones de la función $f(t, y)$ sobre cada intervalo $[t_i, t_{i+1}]$. Estos

métodos resultan de modificar los métodos de Taylor para que el orden de las cotas del error se conserve, pero eliminando la necesidad de determinar y evaluar derivadas parciales de orden alto. La estrategia consiste en aproximar un método de Taylor mediante un método que sea más fácil de evaluar, lo que podría incrementar el error, pero se hace de manera que el incremento no exceda el orden del error de truncamiento que ya presenta el método de Taylor. Como consecuencia, los nuevos errores no influyen significativamente en los cálculos (véase [7]).

En su forma más general, un método RK puede escribirse de la forma

$$u_{i+1} = u_i + \sum_{j=1}^{s} b_j K_j, \quad i \geq 0,$$

donde

$$K_j = h\, f\left(t_i + c_j h,\ u_i + \sum_{k=1}^{s} a_{jk} K_k\right), \quad j = 1, 2 \ldots, s,$$

y s indica el número de evaluaciones de f que hay que efectuar (número de *etapas* del método). Los coeficientes $\{a_{jk}\}$, $\{c_j\}$ y $\{b_j\}$ caracterizan totalmente un método RK y habitualmente se recogen en la llamada *tabla de Butcher*

$$\begin{array}{c|cccc} c_1 & a_{11} & a_{12} & \cdots & a_{1s} \\ c_2 & a_{21} & a_{22} & \cdots & a_{2s} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_s & a_{s1} & a_{s2} & \cdots & a_{ss} \\ \hline & b_1 & b_2 & \cdots & b_s \end{array} \qquad \text{o} \qquad \begin{array}{c|c} \mathbf{c} & A \\ \hline & \mathbf{b}^T \end{array}$$

donde $A = (a_{jk}) \in \mathbb{R}^{s\times s}$, $\mathbf{b} = (b_1, b_2, \ldots, b_s)^T \in \mathbb{R}^s$ y $\mathbf{c} = (c_1, c_2, \ldots, c_s)^T \in \mathbb{R}^s$. Si los coeficientes a_{jk} de A son iguales a cero para $k \geq j$, con $j = 1, 2, \ldots, s$, entonces cada K_j puede calcularse explícitamente en términos de los $j-1$ coeficientes $K_1, K_2, \ldots, K_{j-1}$ que ya han sido determinados. En tal caso el método RK es *explícito*. En caso contrario es *implícito*, siendo necesario resolver un sistema no lineal de tamaño s para calcular los coeficientes.

Uno de los métodos de Runge-Kutta más utilizados es el **método de Runge-Kutta de cuarto orden** *clásico*:

$$u_0 = y_0, \quad u_{i+1} = u_i + \frac{1}{6}(K_1 + 2K_2 + 2K_3 + K_4), \quad i = 0, 1, \ldots, N-1,$$

donde

$$\begin{aligned} K_1 &= h\, f(t_i, u_i), \\ K_2 &= h\, f(t_i + \tfrac{1}{2}h, u_i + \tfrac{1}{2}K_1), \\ K_3 &= h\, f(t_i + \tfrac{1}{2}h, u_i + \tfrac{1}{2}K_2), \\ K_4 &= h\, f(t_i + h, u_i + K_3), \end{aligned} \qquad \begin{array}{c|cccc} 0 & & & & \\ \frac{1}{2} & \frac{1}{2} & & & \\ \frac{1}{2} & 0 & \frac{1}{2} & & \\ 1 & 0 & 0 & 1 & \\ \hline & \frac{1}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{6} \end{array}$$

Este método (que llamaremos RK4) simula la precisión del método de Taylor de orden cuatro y puede deducirse emparejando los coeficientes anteriores con los del método de Taylor de orden cuatro de manera que el error local sea de orden $\mathcal{O}(h^5)$. Es explícito, con un error global de orden $\mathcal{O}(h^4)$ y requiere de cuatro nuevas evaluaciones de f en cada paso.

El **método de Runge-Kutta de segundo orden** (que llamaremos RK2) simula la precisión del método de Taylor de orden dos. Aunque no es un método tan bueno como el RK4, los razonamientos que nos conducen a su desarrollo son más fáciles de entender y sirven para ilustrar las ideas que están involucradas en los métodos de Runge-Kutta.

La forma de la fórmula para el método RK2 se obtiene reemplazando la función $hT_i^{(n)}$ en (6.2) por la función $aK_1 + bK_2$; es decir,

$$u_{i+1} = u_i + aK_1 + bK_2, \tag{6.3}$$

donde

$$K_1 = h\,f(t_i, u_i), \quad K_2 = h\,f(t_i + \alpha h, u_i + \beta K_1),$$

y a, b, α y β son constantes a determinar, de manera que (6.3) sea tan exacta como sea posible.

Para determinar estas constantes, hacemos que la ecuación (6.3) coincida con el desarrollo en serie de Taylor de $y(t)$ en t_i. Por una parte, como

$$\begin{aligned} y(t_{i+1}) &= y(t_i) + hy'(t_i) + \tfrac{h^2}{2}y''(t_i) + \cdots \\ &= y(t_i) + hf(t_i, y(t_i)) + \tfrac{h^2}{2}f'(t_i, y(t_i)) + \cdots \end{aligned}$$

y $f'(t_i, y(t_i)) = f_t(t_i, y(t_i)) + f(t_i, y(t_i))f_y(t_i, y(t_i))$, se sigue que

$$y(t_{i+1}) = y(t_i) + hf(t_i, y(t_i)) + \frac{h^2}{2}\left(f_t(t_i, y(t_i)) + f(t_i, y(t_i))f_y(t_i, y(t_i))\right) + \mathcal{O}(h^3). \tag{6.4}$$

Por otra parte, si ahora desarrollamos $f(t_i + \alpha h, u_i + \beta K_1)$ en serie de Taylor de orden dos para una función de dos variables, se obtiene

$$f(t_i + \alpha h, u_i + \beta K_1) = f(t_i, u_i) + \alpha h f_t(t_i, u_i) + \beta K_1 f_y(t_i, u_i) + \mathcal{O}(h^2),$$

de manera que al sustituir la última expresión en (6.3), queda

$$\begin{aligned} u_{i+1} &= u_i + ahf(t_i, u_i) + bhf(t_i + \alpha h, u_i + \beta K_1) \\ &= u_i + ahf(t_i, u_i) + bh\left(f(t_i, u_i) + \alpha h f_t(t_i, u_i) + \beta K_1 f_y(t_i, u_i)\right) + \mathcal{O}(h^3) \\ &= u_i + (a+b)hf(t_i, u_i) + bh^2\left(\alpha f_t(t_i, u_i) + \beta f(t_i, u_i)f_y(t_i, u_i)\right) + \mathcal{O}(h^3) \end{aligned}$$

Finalmente, si comparamos (6.4) y la expresión anterior, obtenemos el sistema

$$a + b = 1, \qquad \alpha = \beta = \frac{1}{2b},$$

que es un sistema de tres ecuaciones con cuatro incógnitas, pudiéndose elegir entonces una variable arbitrariamente.

Existen por tanto una infinidad de fórmulas de Runge-Kutta de segundo orden, que simulan la precisión del método de Taylor de orden dos, de manera que tendrán un error local de $\mathcal{O}(h^3)$ y un error global de $\mathcal{O}(h^2)$. Por lo tanto, cada fórmula da los mismos resultados exactos si la EDO es cuadrática, lineal o constante. A continuación, damos las tres fórmulas más conocidas.

Si $a = 0$, $b = 1$ y $\alpha = \beta = \frac{1}{2}$, el método RK2 que obtenemos se conoce como **método del punto medio**:

$$u_0 = y_0, \quad u_{i+1} = u_i + K_2, \quad i = 0, 1, \ldots, N-1,$$

donde

$$\begin{aligned} K_1 &= h\,f(t_i, u_i), \\ K_2 &= h\,f(t_i + \tfrac{1}{2}h, u_i + \tfrac{1}{2}K_1), \end{aligned}$$

$$\begin{array}{c|cc} 0 & & \\ \frac{1}{2} & \frac{1}{2} & \\ \hline & 0 & 1 \end{array}$$

Si $a = b = \frac{1}{2}$ y $\alpha = \beta = 1$, el método RK2 que obtenemos se conoce como **método de Euler modificado** (o **método del trapecio**):

$$u_0 = y_0, \quad u_{i+1} = u_i + \frac{1}{2}(K_1 + K_2), \quad i = 0, 1, \ldots, N-1,$$

donde

$$K_1 = h\,f(t_i, u_i), \qquad K_2 = h\,f(t_i + h, u_i + K_1), \qquad \begin{array}{c|cc} 0 & & \\ 1 & 1 & \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$$

Si $a = \frac{1}{4}$, $b = \frac{3}{4}$ y $\alpha = \beta = \frac{2}{3}$, el correspondiente método RK2 se conoce como **método de Ralston** (o **método óptimo**):

$$u_0 = y_0, \quad u_{i+1} = u_i + \frac{1}{4}K_1 + \frac{3}{4}K_2, \quad i = 0, 1, \ldots, N-1,$$

donde

$$K_1 = h\,f(t_i, u_i), \qquad K_2 = h\,f(t_i + \tfrac{2}{3}h, u_i + \tfrac{2}{3}K_1), \qquad \begin{array}{c|cc} 0 & & \\ \frac{2}{3} & \frac{2}{3} & \\ \hline & \frac{1}{4} & \frac{3}{4} \end{array}$$

que corresponde al método RK de segundo orden que tiene un mínimo en el error de truncamiento ([8]).

EJEMPLO. Utilizamos ahora el método del punto medio con $N = 10$ para obtener una aproximación a la solución del problema de valor inicial anterior

$$y'(t) = 2t - y(t), \quad y(0) = -1,$$

en $t = 1$.

Con $h = \frac{1}{10} = 0.1$ y $f(t, y) = 2t - y$, el primer paso del método del punto medio para aproximar $y(0.1)$ es:

$$K_1 = h\,f(t_0, u_0) = (0.1)f(0, -1) = 0.1,$$
$$K_2 = h\,f(t_0 + \tfrac{1}{2}h, u_0 + \tfrac{1}{2}K_1) = (0.1)f(0 + \tfrac{1}{2}(0.1), -1 + \tfrac{1}{2}(0.1)) = 0.105,$$
$$u_1 = u_0 + K_2 = -1 + 0.105 = -0.895.$$

Continuando con las aproximaciones numéricas, se obtiene que en $t = 1$, el valor dado por el método del punto medio es $u_{10} = 0.368541$ y el error absoluto es $|y(1) - u_{10}| = 0.000662$, como puede verse en [16]. Finalizamos la aplicación del método del punto medio mostrando las gráficas de las soluciones exacta y aproximada del PVI anterior en la figura 6.3. □

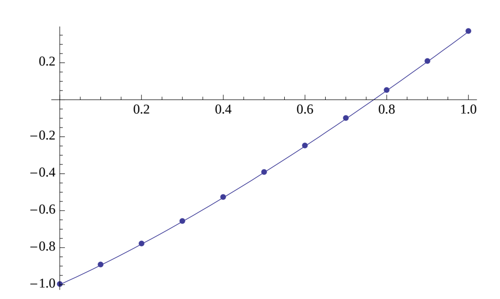


Figura 6.3: Soluciones exacta $y(t) = \mathrm{e}^{-t} + 2t - 2$ (línea continua) y aproximada mediante el método del punto medio con $N = 10$ (puntos).

COMENTARIO ADICIONAL. ▷ Una forma de comparar los métodos de Runge-Kutta de orden bajo es la siguiente. El método RK4 requiere cuatro evaluaciones de f por paso, de manera que para que fuese superior al método de Euler, que requiere solo una evaluación de f por paso, debería dar respuestas más precisas que el método de Euler con tamaño de paso igual a la cuarta parte del tamaño de paso del método RK4. Análogamente, para que el método RK4 fuese superior a los métodos de Runge-Kutta de segundo orden, que requieren dos evaluaciones por paso, debería ser más preciso con tamaño de paso h que un método de segundo orden con tamaño de paso $\frac{h}{2}$. Véase el ejemplo de la pág. 209 de [7] en el que se ilustra la superioridad del método RK4 con respecto a esta medida.

6.5. Métodos multipaso

Los métodos de Taylor y de Runge-Kutta son ejemplos de métodos de un paso para aproximar soluciones de PVI. Estos métodos solo emplean u_i para calcular la aproximación u_{i+1} de $y(t_{i+1})$, sin utilizar explícitamente las aproximaciones previas $u_0, u_1, \ldots, u_{i-1}$. Generalmente requieren de algunas evaluaciones de la función f en puntos intermedios, pero éstas se descartan tan pronto como se obtiene u_{i+1}.

Como la exactitud de $|y(t_j) - u_j|$ disminuye conforme aumenta j, podemos construir mejores métodos de aproximación si, al aproximar $y(t_{i+1})$, incluimos en el método algunas aproximaciones previas u_i. Los métodos que desarrollan esta idea se llaman **métodos de varios pasos** o **métodos multipaso**. Es decir, los métodos de un paso solo tienen en cuenta lo que ocurrió en el paso anterior, mientras que los métodos multipaso tienen en cuenta lo ocurrido en varios pasos anteriores. El principio que los guía se describe a continuación. Deseamos resolver el PVI (6.1). Integrando la EDO de (6.1) entre t_i y t_{i+1}, obtenemos:

$$y(t_{i+1}) - y(t_i) = \int_{t_i}^{t_{i+1}} y'(t)\,dt = \int_{t_i}^{t_{i+1}} f(t, y(t))\,dt \quad \Rightarrow \quad y(t_{i+1}) = y(t_i) + \int_{t_i}^{t_{i+1}} f(t, y(t))\,dt.$$

Puesto que no podemos integrar $f(t, y(t))$ sin conocer $y(t)$, que es la solución del problema, integramos en su lugar el polinomio de interpolación $p(t)$ que aproxima $f(t, y(t))$. Suponiendo además que $y(t_i) \simeq u_i$, tenemos:

$$y(t_{i+1}) \simeq u_i + \int_{t_i}^{t_{i+1}} p(t)\,dt. \tag{6.5}$$

Si u_{m+1} es la primera aproximación que se va a generar usando un método multipaso, entonces necesitamos conocer los valores de partida $u_0, u_1, \ldots, u_m$ del método. Estos valores de partida se generan usando un método de Runge-Kutta o alguna otra técnica de un paso que tenga el mismo orden de error que el método multipaso.

Hay dos clases distintas de métodos multipaso: explícitos e implícitos. En un **método explícito**, el cálculo de u_{i+1} no supone la evaluación de la función $f(t_{i+1}, u_{i+1})$, mientras que en un **método implícito** si.

La precisión de una solución numérica de un PVI está en gran medida determinada por el orden del método utilizado. El orden indica cuantos términos de una solución expresada en serie de Taylor se están simulando mediante este método.

6.5.1. Métodos explícitos de Adams-Bashforth

Supongamos que la fórmula resultante de (6.5) es del tipo

$$u_{i+1} = u_i + c_1 f(t_i, u_i) + c_2 f(t_{i-1}, u_{i-1}) + c_3 f(t_{i-2}, u_{i-2}) + \cdots \tag{6.6}$$

donde $c_1, c_2, c_3, \ldots$ son constantes. Una expresión de este tipo se conoce como *fórmula de Adams-Bashforth* (abreviadamente, AB). A partir del polinomio de interpolación de grado $m-1$, se obtiene una fórmula AB de m pasos.

En las fórmulas de Adams-Bashforth, suponemos que $p(t)$ está dado por el polinomio de interpolación en los m puntos:

$$(t_i, f(t_i, u_i)), (t_{i-1}, f(t_{i-1}, u_{i-1})), \ldots, (t_{i-m+1}, f(t_{i-m+1}, u_{i-m+1})).$$

Si $m = 1$, la correspondiente fórmula AB se reduce al método de Euler. Damos a continuación algunas otras fórmulas, junto con los valores de partida que se requieren y sus términos de error local.

Método explícito de Adams-Bashforth de dos pasos (AB2) ($m = 2$):

$$u_0 = y_0, \quad u_1 = y_1, \quad u_{i+1} = u_i + \frac{h}{2}\left(3f(t_i, u_i) - f(t_{i-1}, u_{i-1})\right),$$

para $i = 1, 2, \ldots, N-1$, con error local $\frac{5}{12} y'''(\xi_i) h^3$, para algún $\xi_i \in (t_{i-1}, t_{i+1})$.

Método explícito de Adams-Bashforth de tres pasos (AB3) ($m = 3$):

$$u_0 = y_0, \quad u_1 = y_1, \quad u_2 = y_2, \quad u_{i+1} = u_i + \frac{h}{12}\left(23f(t_i, u_i) - 16f(t_{i-1}, u_{i-1}) + 5f(t_{i-2}, u_{i-2})\right),$$

para $i = 2, 3, \ldots, N-1$, con error local $\frac{3}{8}y^{(iv)}(\xi_i)h^4$, para algún $\xi_i \in (t_{i-2}, t_{i+1})$.

Método explícito de Adams-Bashforth de cuatro pasos (AB4) ($m = 4$):

$$u_0 = y_0, \quad u_1 = y_1, \quad u_2 = y_2, \quad u_3 = y_3,$$

$$u_{i+1} = u_i + \frac{h}{24}\left(55f(t_i, u_i) - 59f(t_{i-1}, u_{i-1}) + 37f(t_{i-2}, u_{i-2}) - 9f(t_{i-3}, u_{i-3})\right),$$

para $i = 3, 4, \ldots, N-1$, con error local $\frac{251}{720}y^{(v)}(\xi_i)h^5$, para algún $\xi_i \in (t_{i-3}, t_{i+1})$.

Ejemplo. Aplicamos a continuación el método AB4 con $N = 10$ para obtener una aproximación del problema de valor inicial anterior

$$y'(t) = 2t - y(t), \quad y(0) = -1,$$

en $t = 1$.

Podemos obtener los valores iniciales mediante el método RK4, pero como conocemos la solución exacta, $y(t) = \mathrm{e}^{-t} + 2t - 2$, la utilizamos para obtener los valores iniciales:

$$u_0 = -1,\ u_1 = y(0.1) = -0.895162,\ u_2 = y(0.2) = -0.781269,\ u_3 = y(0.3) = -0.659181.$$

Ahora,

$$\begin{aligned} u_4 &= u_3 + \tfrac{h}{24}\left(55f(t_3, u_3) - 59f(t_2, u_2) + 37f(t_1, u_1) - 9f(t_0, u_0)\right) \\ &= -0.659181 + \tfrac{0.1}{24}\left(55f(0.3, -0.659181) - 59f(0.2, -0.781269)\right. \\ &\quad \left. +37f(0.1, -0.895162) - 9f(0, -1)\right) \\ &= -0.659181 + \tfrac{0.1}{24}\left(55(1.259181) - 59(1.181269) + 37(1.095162) - 9(1)\right) \\ &= -0.529677 \end{aligned}$$

y, continuando de forma análoga, véase [16], se tiene: $u_{10} = 0.369344$ y el error cometido es $|y(1) - u_{10}| = 0.00146$. □

Comentario adicional. ▷ La principal desventaja común de todos los métodos multipaso, en general, y de los métodos AB, en particular, es que se necesitan conocer los valores de partida. Por ejemplo, para el método AB4, se necesitan cuatro valores de partida antes de que el método se pueda utilizar. En la práctica, como se ha dicho anteriormente, se utiliza un método de un paso con el mismo orden de error para determinar los valores de partida. El método AB4 se utiliza frecuentemente con el método RK4 porque ambos tienen un error local de orden $\mathcal{O}(h^5)$. La ventaja de los métodos AB sobre los métodos de un paso es que el cálculo de u_{i+1} requiere solo una evaluación de $f(t, y)$ por paso, mientras que los métodos RK de órdenes ≥ 3 requieren cuatro o más evaluaciones de la función. Por esta razón, los métodos multipaso pueden ser el doble de rápidos que los métodos RK de exactitud comparable.

6.5.2. Métodos implícitos de Adams-Moulton

Raramente se utilizan las fórmulas AB de manera aislada, habitualmente se suelen utilizar junto con otras fórmulas para aumentar la precisión de la aproximación numérica. Con el fin de observar cómo es posible esto, volvemos a la aproximación (6.5) y supongamos que se emplea una fórmula de cuadratura numérica que comprende a $f(t_{i+1}, u_{i+1})$. Supongamos ahora que (6.5) da como resultado:

$$u_{i+1} = u_i + c_1 f(t_{i+1}, u_{i+1}) + c_2 f(t_i, u_i) + c_3 f(t_{i-1}, u_{i-1}) + \cdots$$

donde $c_1, c_2, c_3, \ldots$ son constantes. La situación más sencilla es la que considera el polinomio $p(t)$ como el polinomio de interpolación en los $m + 1$ puntos:

$$(t_{i+1}, f(t_{i+1} u_{i+1})), (t_i, f(t_i, u_i)), \ldots, (t_{i-m+1}, f(t_{i-m+1}, u_{i-m+1})).$$

Se relacionan a continuación los ejemplos más comunes de expresiones de este tipo, que se conocen como *fórmulas de Adams-Moulton* (abreviadamente, AM). Obsérvese que el error local de un método implícito de $(m-1)$ pasos es de orden $\mathcal{O}(h^{m+1})$, lo mismo que el de un método explícito de m pasos. Sin embargo, ambos

usan m evaluaciones de la función, ya que los métodos implícitos incluyen $f(t_{i+1}, u_{i+1})$, pero los explícitos no.

Método implícito de Adams-Moulton de dos pasos (AM3) ($m = 2$):

$$u_0 = y_0, \quad u_1 = y_1, \quad u_{i+1} = u_i + \frac{h}{12}\left(5f(t_{i+1}, u_{i+1}) + 8f(t_i, u_i) - f(t_{i-1}, u_{i-1})\right),$$

para $i = 1, 2, \ldots, N-1$, con error local $\frac{-1}{24} y^{(iv)}(\xi_i)h^4$, para algún $\xi_i \in (t_{i-1}, t_{i+1})$.

Método implícito de Adams-Moulton de tres pasos (AM4) ($m = 3$):

$$u_0 = y_0, \quad u_1 = y_1, \quad u_2 = y_2,$$

$$u_{i+1} = u_i + \frac{h}{24}\left(9f(t_{i+1}, u_{i+1}) + 19f(t_i, u_i) - 5f(t_{i-1}, u_{i-1}) + f(t_{i-2}, u_{i-2})\right),$$

para $i = 2, 3, \ldots, N-1$, con error local $\frac{-19}{720} y^{(v)}(\xi_i)h^5$, para algún $\xi_i \in (t_{i-2}, t_{i+1})$.

COMENTARIO ADICIONAL. ▷ Resulta de interés comparar el método explícito AB de m pasos con el método implícito AM de $m-1$ pasos. Ambos requieren m evaluaciones de f por paso y ambos tienen los factores $y^{(m+1)}(\xi_i)h^{m+1}$ en sus términos de error local. Además, en general, los coeficientes de las evaluaciones de f en la fórmula de aproximación y en el término de error local son menores en el método implícito que en el explícito del mismo orden, lo que conlleva menores errores de truncamiento y de redondeo en los métodos implícitos. Por el contrario, obsérvese que los métodos implícitos tienen la necesidad de reformular algebraicamente la ecuación en diferencias para obtener una representación explícita de u_{i+1}, lo que puede ser difícil, sino imposible.

6.5.3. Métodos predictor-corrector

Para sacar provecho de las propiedades beneficiosas de los métodos implícitos y evitar las dificultades en la resolución de la ecuación implícita, estos métodos implícitos no se suelen utilizar en la práctica en solitario, sino que se suelen utilizar para mejorar las aproximaciones obtenidas por los métodos explícitos. La combinación de un método explícito con uno implícito recibe el nombre de **método de predicción y corrección** o **método predictor-corrector**. El método explícito predice una aproximación y el método implícito la corrige. Gozan del orden de precisión del método corrector.

El método predictor es una fórmula explícita y se utiliza en primer lugar para determinar una estimación u_{i+1} de la solución, que se calcula a partir de la solución conocida en el punto anterior mediante un método de un paso o a partir de la solución conocida en varios puntos anteriores mediante un método multipaso. Una vez calculada la estimación u_{i+1}, se aplica el método corrector, que utiliza el valor estimado u_{i+1} en la parte derecha de la ecuación de un método implícito, para calcular una nueva estimación de la solución, más exacta que u_{i+1}, en la parte izquierda de la ecuación del método. Por lo tanto, el método corrector que suele ser un método implícito se utiliza de manera explícita, ya que no se requiere resolver una ecuación no lineal. (Además, podemos repetir la aplicación del método corrector varias veces, de manera que el nuevo valor de u_{i+1} se sustituye de nuevo en la parte derecha de la ecuación del método corrector para obtener una nueva aproximación de u_{i+1} más refinada, véase [12].)

Un método predictor-corrector popular, llamado ABM4, utiliza la fórmula AB4 como método predictor y la fórmula AM4 como método corrector. El primer paso es calcular los valores de partida u_0, u_1, u_2 y u_3 para el método explícito AB4. Para calcular estos cuatro valores, utilizamos un método de un paso de cuarto orden, especificamente, el método RK4. El siguiente paso es calcular una aproximación u_4^P de $y(t_4)$ mediante el método explícito AB4 para hacer la predicción (véase [7]):

$$u_4^P = u_3 + \frac{h}{24}\left(55f(t_3, u_3) - 59f(t_2, u_2) + 37f(t_1, u_1) - 9f(t_0, u_0)\right).$$

Esta aproximación se mejora utilizando ahora el método implícito AM4 como corrector:

$$u_4 = u_3 + \frac{h}{24}\left(9f(t_4, u_4^P) + 19f(t_3, u_3) - 5f(t_2, u_2) + f(t_1, u_1)\right).$$

El valor u_4 es el que se usa como aproximación de $y(t_4)$. Ahora la técnica de utilizar el método AB4 para hacer la predicción y el de AM4 para hacer la corrección se repite, hallándose u_5^P y u_5, las aproximaciones inicial y final respectivas de $y(t_5)$. El proceso se reitera hasta que obtengamos la aproximación de $y(t_N) = y(b)$.

6.6. Ecuaciones de orden superior y sistemas de ecuaciones diferenciales

Los métodos desarrollados en las secciones previas se pueden extender ahora fácilmente para resolver PVI de orden superior. Consideramos el caso general de una EDO de orden $q \geq 2$

$$y^{(q)}(t) = f(t, y(t), y'(t), \ldots, y^{(q-1)}(t)), \text{ para todo } t \in [a,b], \tag{6.7}$$

sujeta a las condiciones iniciales

$$y(a) = y_0, \ y'(a) = y_1, \ \ldots, \ y^{(q-1)}(a) = y_{q-1}.$$

La ecuación (6.7) se puede transformar en un sistema de primer orden de m ecuaciones diferenciales. Para ello, ponemos

$$w_1(t) = y(t), \ w_2(t) = y'(t), \ \ldots, \ w_q(t) = y^{(q-1)}(t),$$

de manera que la ecuación (6.7) puede ahora escribirse como

$$\left\{ \begin{array}{rcl} w_1' & = & w_2, \\ w_2' & = & w_3, \\ & \vdots & \\ w_{q-1}' & = & w_q, \\ w_q' & = & f(t, w_1, w_2, \ldots, w_q), \end{array} \right.$$

con condiciones iniciales

$$w_1(a) = y_0, \ w_2(a) = y_1, \ \ldots, \ w_{q-1}(a) = y_{q-1}.$$

Para resolver el sistema anterior se puede aplicar a cada ecuación individual uno de los métodos desarrollados en las secciones previas.

Ejemplo. Consideramos el método de Euler aplicado a las siguientes dos ecuaciones simultáneas

$$\left\{ \begin{array}{rcl} y_1' & = & f_1(t, y_1(t), y_2(t)), \\ y_2' & = & f_2(t, y_1(t), y_2(t)), \end{array} \right.$$

De la fórmula del método de Euler se sigue que el paso i-ésimo sería

$$\left\{ \begin{array}{rcl} u_{i+1}^1 & = & u_i^1 + h f_1(t_i, u_i^1, u_i^2), \\ u_{i+1}^2 & = & u_i^2 + h f_2(t_i, u_i^1, u_i^2). \quad \square \end{array} \right.$$

Escribiendo el sistema anterior en forma vectorial $\mathbf{y}'(t) = \mathbf{F}(t, \mathbf{y}(t))$, con una elección obvia de la notación, la extensión de los métodos desarrollados anteriormente en el caso de una sola ecuación al caso vectorial es directa. Así, el método

$$\mathbf{u}_0 = \mathbf{y}_0, \quad \mathbf{u}_{i+1} = \mathbf{u}_i + h\,\mathbf{F}(t_i, \mathbf{u}_i), \quad i \geq 0,$$

es la forma vectorial del método de Euler.

6.7. Sugerencias para seguir leyendo

Para un tratamiento más profundo de los métodos numéricos para EDO se recomiendan los siguientes textos: [15] y Ralston y Rabinowitz (2001). Otros textos que ofrecen un estudio exhaustivo son: Butcher (2008), Gear (1971) y Golub y Ortega (1992). Un texto interesante que analiza los sistemas de ecuaciones diferenciales es Lambert (1991), donde se explica que las condiciones para los órdenes de los métodos escalares y vectoriales (sistemas), aunque tengan la «misma forma», varían entre unos y otros, obteniéndose distintos órdenes.

6.8. Ejercicios

1. Aplíquese el método de Euler explícito con paso $h = 0.1$ para aproximar el valor $y(1)$ de la solución de la ecuación integral
$$y(t) = e^t + \int_0^t \cos(s + y(s))\, ds$$
transformándola previamente en un PVI.

2. Encuéntrese la solución exacta del PVI
$$y'(t) = t^2 + y(t), \quad t \in [0, 1], \quad y(0) = 1.$$
Hállese una aproximación numérica obtenida mediante el método de Euler con paso de integración $h = 0.2$.

3. La función $y(t) = \dfrac{t^2}{4}$ es la solución del PVI
$$y'(t) = \sqrt{y(t)}, \quad y(0) = 0.$$
Calcúlese mediante el método de Euler una aproximación de la solución. ¿Qué es lo que sucede? Dese una explicación.

4. *Obtención del método de Euler utilizando integración numérica.* Véase la equivalencia entre aplicar el método de Euler a la resolución del problema de Cauchy (6.1) y aplicar la regla de cuadratura del rectángulo a la reformulación del problema de Cauchy en forma integral.

5. Obténgase el valor exacto de $y(0.1)$ a partir de la solución exacta del PVI
$$y'(t) = -ty(t)^2, \quad y(0) = 2.$$
A continuación, aproxímese numéricamente el valor de $y(0.1)$ mediante el método de Taylor de segundo orden. Compárense los resultados.

6. Constrúyase un método RK explícito de segundo orden y dos etapas tal que $b_1 = \frac{1}{4}$ y dese su tabla de Butcher. Resuélvase el PVI del ejercicio 2 mediante este método con el mismo paso de integración. Compárense los resultados.

7. Si $f(t, y(t))$ depende solo de t (es decir, $f(t, y(t)) = f(t)$) en el problema de Cauchy (6.1), demuéstrese que el método RK2 de Euler modificado se reduce a la regla de cuadratura del trapecio y que el método RK4 se reduce a la regla de cuadratura de Cavalieri-Simpson.

8. Dado el método
$$u_0 = y_0, \quad u_{i+1} = u_i + hf(t_i + h, u_i + hf(t_i, u_i)),$$
dígase si es de tipo RK y, en su caso, escríbase la tabla de Butcher. ¿Cuál es el orden del método?

9. Sea el método
$$\begin{aligned} u_0 = y_0, \quad u_{i+1} &= u_i + \frac{1}{6}(K_1 + 4K_2 + K_3), \\ K_1 &= hf(t_i, u_i), \\ K_2 &= hf(t_i + \tfrac{h}{2}, u_i + \tfrac{1}{2}K_1), \\ K_3 &= hf(t_i + h, u_i + 2K_2 - K_1). \end{aligned}$$
¿Es un método RK? En caso afirmativo, dese su tabla de Butcher. ¿Cuál es el orden del método?

10. Obténgase la fórmula recursiva del método RK4. Para ello, téngase en cuenta que el método RK4 consiste en calcular la aproximación u_{i+1} de la siguiente forma:
$$u_0 = y_0, \quad u_{i+1} = u_i + b_1K_1 + b_2K_2 + b_3K_3 + b_4K_4,$$

donde

$$K_1 = hf(t_i, u_i), \qquad K_3 = hf(t_i + c_3h, u_i + a_{31}K_1 + a_{32}K_2),$$

$$K_2 = hf(t_i + c_2h, u_i + a_{21}K_1), \qquad K_4 = hf(t_i + c_4h, u_i + a_{41}K_1 + a_{42}K_2 + a_{43}K_3).$$

Compruébese que el método es de cuarto orden.

11. Sea el PVI

$$y'(t) = t^2 + y(t), \quad y(0) = 1.$$

Aproxímese el valor $y(1)$ utilizando el método AB2 y tomando como método de inicio el método RK2 dado por

$$u_0 = y_0, \quad u_{i+1} = u_i + hf\left(t_i + \frac{h}{2}, u_i + \frac{h}{2}f(t_i, u_i)\right)$$

con paso de integración $h = 0.2$.

12. *a*) Constrúyanse los métodos AB2 y AM2.
    b) Constrúyanse los métodos AB3 y AM3.

13. Obsérvese el método de Euler explícito y constrúyase de manera análoga el método de Euler implícito, dado por

$$u_0 = y_0, \quad u_{i+1} = u_i + hf(t_{i+1}, u_{i+1}).$$

Utilícense después los dos métodos de Euler en la forma predictor-corrector para resolver el PVI

$$y'(t) = \ln\left(1 + t^2 + y(t)^2\right) + 2t + y(t), \quad y(0) = 1.$$

Escríbanse las ecuaciones del método numérico y calcúlense $y(0.5)$ e $y(1)$ tomando $h = 0.5$.

14. Transfórmese el siguiente PVI de segundo orden

$$y'' + 2y' + y^3 = \operatorname{sen} t, \quad y(0) = 1, \quad y'(0) = 0.$$

en un sistema de ecuaciones diferenciales de primer orden. Calcúlense las dos primeras iteraciones del método de Euler con paso de integración $h = 0.1$.

15. Resuélvase, mediante el método de Euler, el siguiente sistema de ecuaciones diferenciales

$$\begin{cases} x'(t) = x + ty(t) + 1, \\ y'(t) = t^2x(t) + ty(t) + t^3, \end{cases}$$

en el intervalo $[0, 1]$, con las condiciones iniciales $x(0) = 1$, $y(0) = 1$ y usando como paso de integración $h = 0.2$.

16. Desarróllese la fórmula del método RK2 de Euler modificado para la resolución de un sistema de dos EDO de primer orden. Dése también la formulación vectorial del método.

Capítulo 7

Resolución numérica de problemas de valores en la frontera en dos puntos

7.1. Introducción

Las EDO tratadas en el capítulo anterior son de primer orden y deben cumplir una condición inicial. También vimos que las técnicas numéricas se pueden extender a sistemas de ecuaciones y ecuaciones de orden superior, siempre que las condiciones se especifiquen en el mismo punto en el que se da el valor inicial. Por esta razón, tales problemas se denominan problemas de valor inicial (abreviadamente, PVI). Sin embargo, en muchos problemas, las condiciones se especifican en más de un punto. Debido a que estas condiciones se suelen dar en los puntos extremos o frontera de un intervalo, se les denomina *problemas de contorno* o *problemas de valores en la frontera* (abreviadamente, PVF). Muchas aplicaciones importantes de la ingeniería son de esta clase, como por ejemplo la deflexión de una viga y el flujo del calor, así como también varios problemas físicos. Los PVF son generalmente más difíciles de resolver que los PVI. Para EDO de primer orden solo se especifica una condición, así que no hay diferencia entre PVI y PVF.

Presentamos en este capítulo dos procedimientos generales que aproximan la solución de un PVF: los *métodos de disparo* y los *métodos de diferencias finitas*. Los primeros transforman un PVF de segundo orden (u orden superior) en un sistema de PVI. En los segundos, las derivadas de la EDO se aproximan mediante fórmulas de diferencias finitas, resultando entonces sistemas de ecuaciones algebraicas (lineales o no lineales) que hay que resolver. Ambos métodos tienen sus ventajas y desventajas. Los métodos de diferencias finitas no necesitan resolver la EDO varias veces para ajustar las condiciones de contorno prescritas en el punto final del dominio. Por otra parte, la solución de EDO no lineales utilizando métodos de diferencias finitas resulta de la necesidad de resolver un sistema de ecuaciones no lineales simultáneas (generalmente iterativamente), que puede ser tedioso y difícil. Los métodos de disparo tienen la ventaja de que la solución de la EDO no lineal es bastante sencilla. La desventaja de estos métodos es que hay que resolver la EDO varias veces.

7.2. El PVF de segundo orden en dos puntos

Las ecuaciones diferenciales cuyas soluciones aproximaremos aquí son de segundo orden, concretamente de la forma:

$$y''(x) = f(x, y(x), y'(x)), \quad \text{para } x \in [a, b],$$

con las condiciones de contorno que debe cumplir la solución

$$y(a) = \alpha \qquad \text{e} \qquad y(b) = \beta,$$

para ciertas constantes α y β. Estas condiciones de contorno se llaman *condiciones de contorno de Dirichlet.* Un problema como éste es un ejemplo típico de **PVF de segundo orden en dos puntos**.

He aquí un ejemplo de un PVF en dos puntos que se puede resolver mediante funciones elementales:

$$y''(x) + y(x) = 0; \quad y(0) = 1, \quad y(\pi/2) = 2.$$

En primer lugar, podemos encontrar la solución general de la EDO, que es

$$y(x) = C_1 \operatorname{sen} x + C_2 \cos x.$$

Después, podemos determinar las constantes C_1 y C_2 para que las condiciones en la frontera se satisfagan. De este modo,

$$1 = y(0) = C_1 \operatorname{sen} 0 + C_2 \cos 0 = C_2,$$

$$2 = y(\pi/2) = C_1 \operatorname{sen}(\pi/2) + C_2 \cos(\pi/2) = C_1,$$

y la solución es:

$$y(x) = 2 \operatorname{sen} x + \cos x.$$

La técnica que acabamos de ilustrar no es efectiva cuando la solución general de la EDO se desconoce. Nuestro interés se centrará en los métodos numéricos que permiten acometer cualquier PVF en dos puntos. Antes de abordar los métodos numéricos, es conveniente decir que no podemos asegurar que exista una solución del PVF en dos puntos genérico anterior por la sencilla suposición de que f sea una función «decente». En primer lugar, hay que comprobar que se cumplen las siguientes condiciones ([7]):

- la función f y sus derivadas parciales con respecto a y e y' son continuas,
- la derivada parcial de f con respecto a y es positiva,
- la derivada parcial de f con respecto a y' está acotada,

que garantizan que un PVF tiene *solución única* antes de emplear un método numérico; si no se hace, puede que obtengamos resultados absurdos. Éstas son condiciones razonables en los PVF que modelan problemas físicos.

7.3. El método de disparo lineal

Se dice que un PVF es **lineal** cuando la función f es de la forma

$$f(x, y(x), y'(x)) = p(x)y'(x) + q(x)y(x) + r(x),$$

donde $p(x)$, $q(x)$ y $r(x)$ son funciones arbitrarias. Los problemas lineales aparecen frecuentemente en aplicaciones y son mucho más fáciles de resolver que los no lineales. Esto se debe a que sumando una solución particular de la EDO lineal **completa**, o **no homogénea**,

$$y''(x) - p(x)y'(x) - q(x)y(x) = r(x)$$

a la solución general de la EDO lineal homogénea

$$y''(x) - p(x)y'(x) - q(x)y(x) = 0,$$

obtenemos todas las soluciones de la EDO lineal completa. Las soluciones del problema lineal homogéneo son más fáciles de calcular que las del completo. Además, para probar que un PVF lineal tiene *solución única*, solo debemos comprobar que las funciones p, q y r son continuas y que los valores de q son positivos.

Para aproximar la solución única del PVF lineal, consideramos en primer lugar dos PVI

$$y''(x) = p(x)y'(x) + q(x)y(x) + r(x), \quad x \in [a, b]; \quad y(a) = \alpha, \quad y'(a) = 0, \tag{7.1}$$

$$y''(x) = p(x)y'(x) + q(x)y(x), \quad x \in [a, b]; \quad y(a) = 0, \quad y'(a) = 1, \tag{7.2}$$

que tienen solución única. Denotemos la solución del primer problema por $y_1(x)$, la solución del segundo problema por $y_2(x)$ y supongamos que $y_2(b) \neq 0$. Entonces, por la teoría básica de las EDO lineales, es fácil comprobar que la función

$$y(x) = y_1(x) + \frac{\beta - y_1(b)}{y_2(b)} \, y_2(x) \tag{7.3}$$

es la solución única del PVF lineal

$$y''(x) = p(x)y'(x) + q(x)y(x) + r(x), \quad x \in [a, b]; \quad y(a) = \alpha, \quad y(b) = \beta. \tag{7.4}$$

(Los detalles de comprobación se dejan para el estudiante.)

Si por otra parte, $y_2(b) = 0$, entonces la solución $y_2(x)$ de $y''(x) = p(x)y'(x) + q(x)y(x)$ satisface $y_2(a) = y_2(b) = 0$, lo que implica que $y_2(x) \equiv 0$. Por lo tanto, $y_1(x)$ satisface ambas condiciones de contorno.

Así, el método de disparo lineal consiste en sustituir el PVF (7.4) por los dos PVI (7.1) y (7.2), que se resuelven mediante cualquier método numérico de resolución de PVI. Para ello, transformamos sendos problemas en sistemas de primer orden. Para el problema (7.1), el cambio $u_1 = y$, $u_2 = y'$ transforma el PVI (7.1) en el siguiente sistema equivalente

$$\begin{cases} u_1'(x) = u_2(x); & u_1(a) = \alpha, \\ u_2'(x) = q(x)u_1(x) + p(x)u_2(x) + r(x); & u_2(a) = 0, \end{cases}$$

mientras que el cambio $v_1 = y$, $v_2 = y'$ transforma el PVI (7.2) en el sistema equivalente

$$\begin{cases} v_1'(x) = v_2(x); & v_1(a) = 0, \\ v_2'(x) = q(x)v_1(x) + p(x)v_2(x); & v_2(a) = 1. \end{cases}$$

Una vez que se dispone de las aproximaciones $y_1(x)$ e $y_2(x)$, la solución del PVF (7.4) se aproxima usando la expresión (7.3).

Ejemplo. Sea el PVF

$$y''(x) + (x+1)y'(x) - 2y(x) = (1 - x^2)\,\mathrm{e}^{-x}, \quad x \in [0, 1]; \quad y(0) = y(1) = 0.$$

Vamos a resolverlo utilizando el método de disparo lineal con $h = 0.2$.

Las funciones p, q y r son respectivamente $p(x) = -(x + 1)$, $q(x) = 2$ y $r(x) = (1 - x^2)\,\mathrm{e}^{-x}$. La solución numérica del problema anterior es:

$$y(x) = y_1(x) - \frac{y_1(1)}{y_2(1)}\, y_2(x), \tag{7.5}$$

donde $y_1(x)$ e $y_2(x)$ son las soluciones respectivas de los PVI:

$$y_1''(x) = -(x+1)y_1'(x) + 2y_1(x) + (1 - x^2)\,\mathrm{e}^{-x}; \quad y_1(0) = y_1'(0) = 0,$$
$$y_2''(x) = -(x+1)y_2'(x) + 2y_2(x); \quad y_2(0) = 0, \quad y_2'(0) = 1,$$

que los expresamos ahora respectivamente de la forma:

$$\begin{cases} u_1'(x) = u_2(x); & u_1(0) = 0, \\ u_2'(x) = 2u_1(x) - (x+1)u_2(x) + (1 - x^2)\,\mathrm{e}^{-x}; & u_2(0) = 0, \quad x \in [0, 1], \end{cases} \tag{7.6}$$

$$\begin{cases} v_1'(x) = v_2(x); & v_1(0) = 0, \\ v_2'(x) = 2v_1(x) - (x+1)v_2(x); & v_2(0) = 1, \quad x \in [0, 1]. \end{cases} \tag{7.7}$$

Utilizamos el método RK4 para construir las soluciones numéricas u_i^1 y v_i^1 de los sistemas lineales (7.6) y (7.7). Las aproximaciones de u_i^1 y v_i^1 están dadas en la tabla 7.1 (véase [16]).

x	u_i^1	v_i^1
0	0	0
0.2	0.01743428	0.18251267
0.4	0.06017154	0.33919207
0.6	0.11620204	0.48140428
0.8	0.17679396	0.61776325
1	0.23633393	0.75454368

Tabla 7.1: Soluciones numéricas de los sistemas (7.6) y (7.7).

Por ejemplo, utilizando (7.5) y la tabla 7.1, la solución aproximada en $x = 0.4$ es

$$y(0.4) = y_1(0.4) - \frac{y_1(1)}{y_2(1)}\, y_2(0.4) \approx 0.06017154 - \frac{0.23633393}{0.75454368}\,(0.33919207) = -0.046068294. \quad \square$$

Comentario adicional. ▷ En general, si u_i^1 y v_i^1 son aproximaciones respectivas de $y_1(x_i)$ e $y_2(x_i)$ con una precisión de orden $\mathcal{O}(h^n)$, para cada $i = 0, 1, \dots, N$, entonces la aproximación de $y(x_i)$ tendrá también una precisión de orden $\mathcal{O}(h^n)$.

7.4. Métodos de diferencias finitas lineales

El método de disparo lineal suele presentar problemas con los errores de redondeo. Los métodos que presentamos en esta sección son más robustos frente a los errores de redondeo, pero, en general, requieren mayor esfuerzo computacional para obtener una precisión específica.

Los métodos que utilizan diferencias finitas para resolver problemas de contorno sustituyen cada una de las derivadas de la EDO por un cociente incremental como los considerados en el capítulo 4; los cocientes incrementales concretos que se tomen deben mantener un orden del error especificado. Así, los métodos que involucran diferencias finitas transforman el PVF en un sistema lineal o no lineal, cuadrado, cuyas incógnitas serán los valores aproximados de la solución $y(x)$ en los puntos elegidos del intervalo $[a, b]$.

En el **método de diferencias finitas** para el PVF (7.4) es necesario utilizar cocientes incrementales para aproximar tanto y' como y''. Para cualquier x del intervalo (a, b), sencillas manipulaciones del desarrollo de Taylor de $y(x)$, alrededor de x, nos permite obtener la diferencia finita centrada para y' (véase el capítulo 4):

$$y'(x) \simeq \frac{y(x+h) - y(x-h)}{2h}.$$

De forma análoga, para la derivada segunda, obtenemos la diferencia finita centrada para y''

$$y''(x) \simeq \frac{y(x+h) - 2y(x) + y(x-h)}{h^2}.$$

Sustituyendo entonces las diferencias finitas centradas que acabamos de ver en la EDO, resulta

$$\frac{y(x+h) - 2y(x) + y(x-h)}{h^2} = p(x)\,\frac{y(x+h) - y(x-h)}{2h} + q(x)y(x) + r(x). \tag{7.8}$$

Tomando ahora un número entero $N > 0$ y dividiendo el intervalo $[a, b]$ en $(N + 1)$ subintervalos del mismo tamaño h, con

$$h = \frac{b-a}{N+1}, \quad x_i = a + ih, \quad i = 0, 1, \dots, N+1,$$

podemos discretizar la expresión (7.8) para cada x_i, de manera que

$$\frac{y(x_{i+1}) - 2y(x_i) + y(x_{i-1})}{h^2} = p(x_i)\,\frac{y(x_{i+1}) - y(x_{i-1})}{2h} + q(x_i)y(x_i) + r(x_i), \quad i = 1, 2, \dots, N.$$

Si utilizamos la notación $u_i = y(x_i)$, $p_i = p(x_i)$, $q_i = q(x_i)$ y $r_i = r(x_i)$, con $i = 1, 2, \dots, N$, $u_0 = \alpha$ y $u_{N+1} = \beta$, y agrupamos términos, obtenemos el sistema lineal de tamaño $N \times N$

$$\left(1 + \frac{h}{2}p_i\right) u_{i-1} - \left(2 + h^2 q_i\right) u_i + \left(1 - \frac{h}{2}p_i\right) u_{i+1} = h^2 r_i, \quad i = 1, 2, \dots, N, \tag{7.9}$$

que se puede representar en la forma matricial $A\mathbf{u} = \mathbf{b}$, donde A es una matriz $N \times N$ tridiagonal,

$$\begin{pmatrix} b_1 & c_1 & & & & \\ a_2 & b_2 & c_2 & & & \\ & a_3 & b_3 & c_3 & & \\ & & \ddots & \ddots & \ddots & \\ & & & a_{N-1} & b_{N-1} & c_{N-1} \\ & & & & a_N & b_N \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \\ u_{N-1} \\ u_N \end{pmatrix} = \begin{pmatrix} d_1 - a_1\alpha \\ d_2 \\ d_3 \\ \vdots \\ d_{N-1} \\ d_N - c_N\beta \end{pmatrix},$$

donde a_i, b_i, c_i y d_i están definidos, para $i = 1, 2, \ldots, N$, por

$$a_i = 1 + \frac{h}{2}p_i, \qquad b_i = -(2 + h^2 q_i), \qquad c_i = 1 - \frac{h}{2}p_i, \qquad d_i = h^2 r_i.$$

Este sistema tiene *solución única* si p, q y r son continuas en $[a, b]$, $q(x) \geq 0$ en $[a, b]$ y $h < 2/L$, donde $L = \max\limits_{a \leq x \leq b} |p(x)|$.

Ejemplo. Sea el PVF

$$y''(x) + (x + 1)y'(x) - 2y(x) = (1 - x^2)\,\mathrm{e}^{-x}, \quad x \in [0, 1]; \quad y(0) = -1, \quad y(1) = 0.$$

Utilizaremos el método de diferencias finitas lineales con $h = 0.2$ para resolverlo y compararemos los resultados con la solución exacta $y(x) = (x - 1)\,\mathrm{e}^{-x}$.

Las funciones p, q y r son respectivamente $p(x) = -(x + 1)$, $q(x) = 2$ y $r(x) = (1 - x^2)\,\mathrm{e}^{-x}$. Por tanto, la ecuación (7.9) se convierte en

$$(1 - 0.1(x_i + 1))u_{i-1} - (2 + 0.08)u_i + (1 + 0.1(x_i + 1))u_{i+1} = 0.04(1 - x_i^2)\,\mathrm{e}^{-x_i},$$

con $u_0 = -1$, $u_5 = 0$ y $x_i = (0.2)i$, para $i = 1, 2, 3, 4$. Y la correspondiente formulación matricial es

$$\begin{pmatrix} -2.08 & 1.12 & & \\ 0.86 & -2.08 & 1.14 & \\ & 0.84 & -2.08 & 1.16 \\ & & 0.82 & -2.08 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{pmatrix} = \begin{pmatrix} 0.91143926 \\ 0.02252275 \\ 0.01404958 \\ 0.00647034 \end{pmatrix}. \tag{7.10}$$

Las aproximaciones de la solución aparecen en la tabla 7.2 (véase [16]), junto con su comparación con los valores calculados a partir de la solución analítica. □

x	u_i	$y(x_i)$	Error absoluto
0	-1.00000	-1.00000	0.000000
0.2	-0.65413	-0.65498	0.000854
0.4	-0.40103	-0.40219	0.001163
0.6	-0.21848	-0.21952	0.001047
0.8	-0.08924	-0.08987	0.000624
1	0.00000	0.00000	0.000000

Tabla 7.2: Solución numérica del sistema (7.10).

Comentarios adicionales. ▷ De la aplicación de los métodos de diferencias finitas no siempre resulta un sistema tridiagonal de ecuaciones. Esto último es debido a que la EDO es de segundo orden y se ha utilizado una fórmula de diferencias finitas centradas para aproximar la derivada segunda.

▷ Obsérvese que al utilizar diferencias finitas centradas como aproximaciones de las derivadas, resulta que el método de diferencias finitas tendrá un error de truncamiento de orden $\mathcal{O}(h^2)$. Para obtener un método de diferencias más preciso tenemos varias opciones. Podemos utilizar aproximaciones en diferencias finitas más precisas para las derivadas, pero la desventaja de hacer esto es que el sistema de ecuaciones resultante no será tridiagonal y su resolución requerirá entonces muchas más operaciones. Otro camino, generalmente más satisfactorio, consiste en considerar una reducción del tamaño de paso h y comparar las soluciones en los mismos nodos; sin embargo, el error de redondeo incrementará y podría llegar a hacerse grande.

7.5. El método de disparo no lineal

Los métodos de diferencias finitas trabajan razonablemente bien para PVF lineales y no presentan probemas de inestabilidad. Para PVF con EDO no lineales, estos métodos suelen presentar problemas como consecuencia de que el sistema resultante es no lineal. En tales situaciones es preferible utilizar el **método de disparo no lineal**.

La técnica del disparo para un PVF no lineal

$$y''(x) = f(x, y(x), y'(x)), \quad x \in [a, b]; \quad y(a) = \alpha, \quad y(b) = \beta, \tag{7.11}$$

es parecida a la del método de disparo lineal, salvo que la solución de un problema no lineal no puede expresarse como una combinación lineal de soluciones de dos PVI. En su lugar, aproximamos la solución del PVF mediante las soluciones de una sucesión de PVI de la forma

$$y''(x) = f(x, y(x), y'(x)), \quad x \in [a, b]; \quad y(a) = \alpha, \quad y'(a) = m, \tag{7.12}$$

que involucran un cierto parámetro m, que indicará la pendiente de salida de la solución del problema (de ahí el nombre de método de *disparo* para esta técnica, que viene de la analogía con el procedimiento para disparar contra un blanco fijo, véase [7]). Para ello, elegiremos valores m_k del parámetro m de manera que se garantice

$$\lim_{k\to\infty} y(b; m_k) = y(b) = \beta,$$

donde $y(x; m_k)$ denota la solución del PVI (7.12) con $m = m_k$ e $y(x)$ la solución del PVF (7.11). Observemos que $y(x; m_k)$ satisface todas las condiciones para ser la solución del PVF (7.11), a falta de que su valor en $x = b$ sea β.

Iniciamos el proceso con un valor del parámetro $m = m_0$ que determina la elevación con la que se dispara desde el punto (a, α), de manera que la trayectoria de la bala viene dada por la solución del PVI

$$y''(x) = f(x, y(x), y'(x)), \quad x \in [a, b]; \quad y(a) = \alpha, \quad y'(a) = m_0.$$

Si $y(b; m_0)$ no está suficientemente cerca de β, corregimos nuestra aproximación tomando una nueva elevación $m_1, m_2, \ldots$, y así sucesivamente hasta que $y(b; m_k)$ «acierte» en β con la precisión deseada.

El problema es entonces determinar el parámetro m del PVI de manera que

$$y(b; m) - \beta = 0$$

(es decir, se trata de resolver la ecuación anterior en la variable m). Por tanto, todo se reduce a resolver una ecuación no lineal, descrita por una función de la que solo conocemos su valor en una serie de puntos, para lo cual se puede aplicar cualquier método numérico de los descritos en el capítulo 2. Por ejemplo, podemos recurrir al método de bisección, al de la secante o al de Newton. Pero el empleo de un método u otro necesitará de un mayor o menor conocimiento de la solución de (7.12) y, quizá, de alguna derivada, lo cual puede hacer complicada su aplicación.

Un método sencillo que se puede utilizar para determinar el parámetro m es el *método de la secante*, para el que se necesita dar dos valores iniciales m_0 y m_1 y hallar las siguientes aproximaciones mediante la iteración

$$m_k = m_{k-1} - (y(b; m_{k-1}) - \beta)\, \frac{m_{k-1} - m_{k-2}}{y(b; m_{k-1}) - y(b; m_{k-2})}, \quad k = 2, 3 \ldots \tag{7.13}$$

A continuación, conocido el valor de m_k en la k-ésima iteración, se vuelve a resolver el PVI (7.12) para $m = m_k$ y, mediante (7.13) con $k = k + 1$, se obtiene un nuevo valor de la pendiente de disparo $m = m_{k+1}$. Y así hasta alcanzar un m_j tal que $|y(b; m_j) - \beta| < \epsilon$, donde ϵ es la cota del error que estemos dispuestos a cometer. Notemos que para que el método converja, son necesarias buenas elecciones de m_0 y m_1. La convergencia del método de la secante, en condiciones adecuadas, asegura la convergencia del método de disparo no lineal.

En la práctica, ninguno de los PVI (7.12) se resuelve exactamente, sino que sus soluciones se aproximan utilizando alguno de los métodos numéricos tratados en el capítulo anterior.

Ejemplo. Consideremos el PVF

$$yy'' = -(y')^2, \quad x \in [1, 3]; \quad y(1) = \sqrt{2}, \quad y(3) = 2,$$

cuya solución exacta es $y = \sqrt{1+x}$. Para este ejemplo, $f(x, y, y') = -(y')^2/y$, $a = 1$, $b = 3$, $\alpha = \sqrt{2}$ y $\beta = 2$.

Para aplicar el método de disparo no lineal para resolver este problema con una precisión de 10^{-6}, podemos tomar como primera aproximación $m_0 = y(1) = \sqrt{2} \approx 1.4$ y reemplazar el PVF por el PVI

$$yy'' = -(y')^2, \quad x \in [1,3]; \quad y(1) = \sqrt{2}, \quad y'(1) = 1.4.$$

Transformamos ahora la EDO anterior en un sistema de primer orden de dos EDO, ajustamos las condiciones iniciales al sistema y aplicamos a cada EDO individual el método RK4 para obtener

$$y(3; m_0) = 3.14953128, \qquad |y(3; m_0) - 2| = 1.14953128.$$

Repetimos el proceso utilizando una estimación diferente de $y'(1)$; por ejemplo, $m_1 = \frac{\beta-\alpha}{b-a} = \frac{2-\sqrt{2}}{2} \approx 0.3$, de manera que

$$y(3; m_1) = 1.92277306, \qquad |y(3; m_1) - 2| = 0.07722694.$$

Para la tercera aproximación utilizamos el método de la secante

$$m_2 = 0.3 - \frac{(1.92277306 - 2)(0.3 - 1.4)}{1.92277306 - 3.14953128} \approx 0.30629520$$

y obtenemos

$$y(3; m_2) = 2.02207265, \qquad |y(3; m_2) - 2| = 0.02207265.$$

Se repite el proceso para calcular las siguientes $y(3; m_k)$. Después de seis intentos, véase [16], se obtiene el valor $m_5 = 0.35355340$ para el que $y(3; m_5) = 2$ hasta 10 cifras decimales. □

COMENTARIO ADICIONAL. ▷ Análogamente, se puede utilizar el método de Newton para hallar una solución aproximada de la ecuación $y(b; m) - \beta = 0$ en cada iteración, pero la necesidad de conocer la derivada de la solución en la iteración anterior lleva consigo tener que resolver un nuevo PVI, aunque generalmente converge más rápidamente que utilizando el método de la secante. Ambos métodos presentan solo convergencia local, ya que ambos requieren buenas aproximaciones iniciales. (Véanse [5, 7, 8] para la resolución mediante el método de Newton.)

NOTA. Debido a la complejidad de estudio del error en estos métodos, no vamos a indicar nada sobre el mismo. Para un estudio detallado de error, puede consultarse [15].

7.6. Métodos de diferencias finitas no lineales

El método de diferencias finitas para un PVF general (7.11) es parecido al método que se aplica en los problemas lineales, excepto que el sistema de ecuaciones resultante es no lineal, y habrá que recurrir a los métodos numéricos correspondientes para su resolución. Véase [5].

Al igual que en el caso lineal, dividimos el intervalo $[a, b]$ en $N+1$ subintervalos cuyos extremos son los nodos $x_i = a + ih$, para $i = 0, 1, \ldots, N+1$. Suponiendo que la solución exacta tiene derivada cuarta acotada, podemos aproximar $y'(x)$ e $y''(x)$ por las diferencias finitas centradas adecuadas, discretizar la expresión resultante para $x = x_i$, $i = 1, 2, \ldots, N$, y obtener

$$\frac{y(x_{i+1}) - 2y(x_i) + y(x_{i-1})}{h^2} = f\left(x_i, y(x_i), \frac{y(x_{i+1}) - y(x_{i-1})}{2h}\right), \quad i = 1, 2, \ldots, N.$$

Para generar el método de diferencias finitas, suprimimos el término de error y añadimos las condiciones de contorno, lo que transforma el PVF en el sistema no lineal de tamaño $N \times N$

$$\left.\begin{array}{rcl} 2u_1 - u_2 + h^2 f\left(x_1, u_1, \frac{u_2-\alpha}{2h}\right) - \alpha &=& 0 \\ -u_1 + 2u_2 - u_3 + h^2 f\left(x_2, u_2, \frac{u_3-u_1}{2h}\right) &=& 0 \\ &\vdots& \\ -u_{N-2} + 2u_{N-1} - u_N + h^2 f\left(x_{N-1}, u_{N-1}, \frac{u_N-u_{N-2}}{2h}\right) &=& 0 \\ -u_{N-1} + 2u_N + h^2 f\left(x_N, u_N, \frac{\beta-u_{N-1}}{2h}\right) - \beta &=& 0 \end{array}\right\},$$

donde $u_0 = \alpha$, $u_{N+1} = \beta$ y u_i representa el valor aproximado de la solución $y(x)$ en el punto $x = x_i$. Para aproximar la solución del sistema anterior, recurrimos al método de Newton para sistemas no lineales. Se genera entonces una sucesión de iteraciones $\left\{\left(u_1^{(k)}, u_2^{(k)}, \ldots, u_N^{(k)}\right)^T\right\}$ que converge a la solución del sistema suponiendo que la aproximación inicial $\left(u_1^{(0)}, u_2^{(0)}, \ldots, u_N^{(0)}\right)^T$ esté lo suficientemente cerca de la solución $(u_1, u_2, \ldots, u_N)^T$. Si denotamos por $z_i = \dfrac{u_{i+1} - u_{i-1}}{2h}$, $i = 1, 2, \ldots, N$, la matriz jacobiana que aparece en el método de Newton es la matriz tridiagonal

$$J = \begin{pmatrix} b_1 & c_1 & & & & \\ a_2 & b_2 & c_2 & & & \\ & a_3 & b_3 & c_3 & & \\ & & \ddots & \ddots & \ddots & \\ & & & a_{N-1} & b_{N-1} & c_{N-1} \\ & & & & a_N & b_N \end{pmatrix},$$

donde a_i, b_i y c_i están definidos, para $i = 1, 2, \ldots, N$, por

$$a_i = -\left(1 + \frac{h}{2} f_{y'}(x_i, u_i, z_i)\right), \quad b_i = 2 + h^2 f_y(x_i, u_i, z_i), \quad c_i = -\left(1 - \frac{h}{2} f_{y'}(x_i, u_i, z_i)\right).$$

El método de Newton para sistemas no lineales requiere que, en cada iteración, se resuelva un sistema lineal $N \times N$ con la anterior matriz tridiagonal J ([5])

$$J \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_{N-1} \\ v_N \end{pmatrix} = \begin{pmatrix} 2u_1 - u_2 + h^2 f\left(x_1, u_1, \frac{u_2 - \alpha}{2h}\right) - \alpha \\ -u_1 + 2u_2 - u_3 + h^2 f\left(x_2, u_2, \frac{u_3 - u_1}{2h}\right) \\ \vdots \\ -u_{N-2} + 2u_{N-1} - u_N + h^2 f\left(x_{N-1}, u_{N-1}, \frac{u_N - u_{N-2}}{2h}\right) \\ -u_{N-1} + 2u_N + h^2 f\left(x_N, u_N, \frac{\beta - u_{N-1}}{2h}\right) - \beta \end{pmatrix},$$

donde $\mathbf{v} = (v_1, v_2, \ldots, v_N)^T$ nos permite actualizar la iteración del método de Newton

$$\mathbf{u}^{(k)} = \mathbf{u}^{(k-1)} + \mathbf{v},$$

con $\mathbf{u}^{(k)} = \left(u_1^{(k)}, u_2^{(k)}, \ldots, u_N^{(k)}\right)^T$. Como en este caso la matriz J es tridiagonal, podemos aplicar el método de factorización de Crout. Las aproximaciones iniciales $u_i^{(0)}$ de u_i, para cada $i = 1, 2, \ldots, N$, son las ordenadas correspondientes a las abcisas x_i en la línea recta que pasa por los puntos (a, α) y (b, β).

Se puede demostrar, véase [15], que el método anterior es de orden $\mathcal{O}(h^2)$.

Como es posible que necesitemos una buena aproximación inicial, debemos establecer una cota del número de iteraciones y, si se excede, buscar una nueva aproximación o disminuir el tamaño de paso.

7.7. Sugerencias para seguir leyendo

Para la teoría básica se pueden consultar las referencias dadas en el capítulo anterior. También se pueden consultar los siguientes textos: Keller (1992), Fox (1990) o Golub y Ortega (1992).

7.8. Ejercicios

1. Denuéstrese que los siguientes PVF tienen solución única

   a) $y'' = (5y + \operatorname{sen} 3y)\, \mathrm{e}^x; \quad y(0) = y(1) = 0,$

b) $y'' + \mathrm{e}^{-xy} + \operatorname{sen} y' = 0; \quad y(1) = y(2) = 0.$

2. Demuéstrese que el PVF lineal

$$y''(x) = p(x)y'(x) + q(x)y(x) + r(x), \quad x \in [a,b]; \quad y(a) = \alpha, \quad y(b) = \beta,$$

tiene solución única si $p(x)$, $q(x)$ y $r(x)$ son continuas en $[a,b]$ y $q(x) > 0$ en $[a,b]$.

3. Aplíquese, usando el método de Euler con paso $h = 0.5$ para los PVI asociados, el método de disparo lineal para aproximar la solución del PVF

$$y'' = e^x y + (1+x)y' + x^3; \quad y(0) = 0, \quad y(1) = 2.$$

4. Escríbase la ecuación en diferencias lineales con $N = 9$ que aproxima la solución del PVF

$$y'' + 3y' + 2y = 4x^2; \quad y(1) = 1, \quad y(2) = 6.$$

5. Pruébese que el sistema lineal $A\mathbf{u} = \mathbf{b}$, donde A es una matriz $N \times N$ tridiagonal, correspondiente al método de diferencias finitas lineales, tiene solución única si p, q y r son continuas en $[a,b]$, $q(x) \geq 0$ en $[a,b]$ y $h < \frac{2}{L}$, donde $L = \text{máx}_{a \leq x \leq b} |p(x)|$. (*Ayuda:* pruébese primero que A es una matriz estrictamente diagonal dominante.)

6. Supongamos que $p(x) \equiv C_1 > 0$ y $q(x) \equiv C_2 > 0$.

   a) Escríbase el sistema lineal que representa al método de diferencias finitas lineales para esta situación.

   b) Prúebese que el sistema tiene solución única si $\frac{C_1}{C_2} \leq h$.

7. Utilícese el método de diferencias finitas lineales con $h = \frac{1}{4}$ para aproximar la solución del PVF

$$y'' = 4(y - x); \quad y(0) = 0, \quad y(1) = 2.$$

Compárense los resultados con la solución exacta $y(x) = e^2(e^4 - 1)^{-1}(e^{2x} - e^{-2x}) + x$.

8. Si aplicamos el método de disparo no lineal al PVF

$$y'' = \frac{y}{2} - \frac{2}{y}(y')^2; \quad y(0) = 1, \quad y(1) = 1.5,$$

y obtenemos los siguientes resultados

$$y'(0) = 0 \Rightarrow y(1) = 1.54308, \qquad y'(0) = -0.005 \Rightarrow y(1) = 1.42556,$$

¿qué valor de $y'(0)$ debería utilizarse en el siguiente disparo?

9. Sea el PVF

$$y'' = -(y')^2 - y + \ln x; \quad y(1) = 0, \quad y(2) = \ln 2.$$

   a) Disparando con $m_0 = 0$ y $m_1 = 1$, utilícese el método de disparo no lineal (usando como método para los PVI asociados el método de Euler con paso $h = 0.5$) para calcular las dos primeras correcciones al disparo inicial (m_2 y m_3), y aproxímese la solución del PVF con disparo m_3.

   b) Mediante el método de diferencias finitas no lineales con $h = 0.5$, aproxímese la solución del PVF.

   c) Compárense los resultados con la solución exacta $y(x) = \ln x$.

10. Utilícese el método de disparo no lineal con $h = 0.25$ para aproximar la solución del PVF

$$y''(x) = y'(x) + 2(y(x) - \ln x)^3 - \frac{1}{x}; \quad y(2) = \frac{1}{2} + \ln 2, \quad y(3) = \frac{1}{3} + \ln 3,$$

disparando con $m_0 = \frac{\beta - \alpha}{b - a}$ y $m_1 = m_0 + \frac{\beta - y(b;m_0)}{b-a}$. Compárense los resultados con $y(x) = \frac{1}{x} + \ln x$.

11. Aproxímese la solución del PVF

$$y''(x) = 2y(x)^3; \quad y(-1) = \frac{1}{2}, \quad y(0) = \frac{1}{3},$$

mediante el método de diferencias finitas no lineales con $h = 0.25$. Compárense los resultados con la solución exacta $y(x) = \frac{1}{x+3}$.

12. *Condiciones de contorno de Neumann.* Las condiciones de contorno que aparecen en los PVF se pueden especificar de diferentes maneras. Si en vez de ser dos valores de y, son dos valores de y' de la forma $y'(a) = \alpha$ e $y'(b) = \beta$, hablaremos de condiciones de contorno de Neumann. En estos problemas, los métodos de diferencias finitas se utilizan para aproximar la EDO en los puntos interiores. Sin embargo, el sistema algebraico que se obtiene no se puede resolver porque la solución en los puntos extremos no está dada (luego hay más incógnitas que ecuaciones). Las ecuaciones adicionales necesarias para resolver el problema se obtienen aproximando las condiciones de contorno mediante diferencias finitas e incorporando las ecuaciones resultantes a las ecuaciones algebraicas obtenidas para los puntos interiores. Aplíquese lo anterior al PVF

$$y''(x) + \frac{1}{x}\, y'(x) = 1; \quad y'(0) = 0, \quad y'(1) = y(1) + y(1)^4$$

utilizando diferencias finitas centradas y asegurándose de que el orden del error de truncamiento de las condiciones de Neumann es compatible con el de la EDO.

13. *Condiciones de contorno mixtas.* Si las condiciones de contorno son de la forma $y'(a) + c_1 y(a) = \alpha$ e $y'(b) + c_2 y(b) = \beta$, donde c_1 y c_2 son constantes, reciben el nombre de condiciones de contorno mixtas. Aproxímese el PVF

$$y''(x) = y(x) + \mathrm{e}^x; \quad y(0) = 1, \quad y'(1) = -1$$

utilizando diferencias finitas centradas para todas las derivadas y dividiéndose el dominio de integración en cinco subintervalos. Compárense los resultados numéricos con la solución exacta $y(x) = \frac{1-\mathrm{e}-\mathrm{e}^2}{1+\mathrm{e}^2}\,\mathrm{e}^x + \frac{\mathrm{e}(1+2\,\mathrm{e})}{1+\mathrm{e}^2}\,\mathrm{e}^{-x} + \frac{x}{2}\,\mathrm{e}^x$.

14. Si la EDO lineal $y''(x) - p(x)y'(x) - q(x)y(x) = r(x)$ está sujeta a las condiciones contorno $y(a) = \alpha$ e $y'(b) + cy(b) = \beta$, la combinación lineal $y(x) = y_1(x) + \mu y_2(x)$, donde $y_1(x)$ es la solución del problema (7.1) e $y_2(x)$ la del problema (7.2), satisface la condición $y(a) = \alpha$. Si queremos aplicar el método del disparo lineal, necesitamos encontrar μ para que $y(x)$ satisfaga $y'(b) + cy(b) = \beta$. Si $y_2'(b) + cy_2(b) \neq 0$, hay una única solución dada, la dada por

$$y(x) = y_1(x) + \frac{\beta - y_1'(b) - cy_1(b)}{y_2'(b) + cy_2(b)}\, y_2(x).$$

Si por el contrario $y_2'(b) + cy_2(b) = 0$, es fácil ver que $y(x) = y_1(x)$ satisface ambas condiciones de contorno. Aproxímese entonces la solución del PVF del ejercicio anterior utilizando el método de disparo lineal.

15. Para poder aproximar la solución del PVF

$$y''(x) - p(x)y'(x) - q(x)y(x) = r(x); \quad y'(a) + c_1 y(a) = \alpha, \quad y'(b) + c_2 y(b) = \beta,$$

mediante el método del disparo lineal, los dos PVI que hay que resolver son:

$$y''(x) = p(x)y'(x) + q(x)y(x) + r(x), \quad x \in [a, b]; \quad y(a) = 0, \quad y'(a) = \alpha,$$

$$y''(x) = p(x)y'(x) + q(x)y(x), \quad x \in [a, b]; \quad y(a) = 1, \quad y'(a) = -c_1.$$

La combinación lineal $y(x) = y_1(x) + \mu y_2(x)$, donde $y_1(x)$ es la solución del primer PVI e $y_2(x)$ la del segundo, satisface la condición $y'(a) + c_1 y(a) = \alpha$. Necesitamos entonces encontrar, si es posible, μ para que $y(x)$ satisfaga $y'(b) + c_2 y(b) = \beta$. Si $y_2'(b) + c_2 y_2(b) \neq 0$, hay una única solución dada, la dada por

$$y(x) = y_1(x) + \frac{\beta - y_1'(b) - c_2 y_1(b)}{y_2'(b) + c_2 y_2(b)}\, y_2(x).$$

Aplíquese lo anterior para aproximar la solución del PVF

$$y''(x) + y'(x) + y(x) = 5x; \quad y'(0) + 3y(0) = 7, \quad 7y'(1) + y(1) = 8,$$

mediante el método de disparo lineal.

16. Los problemas relacionados con el campo de la elasticidad y la vibración se ubican en una clase especial de PVF denominada *problemas de valores propios*. Algunos ejemplos de este tipo de PVF son los que aparecen a continuación.

    a) $y'' + \lambda y = 0; \quad y(0) = y(1) = 0,$

    b) $y'' + \lambda y = 0; \quad y'(0) = y(1) = 0,$

    c) $y'' - 2y' + (1 + \lambda)y = 0; \quad y(0) = 0, \quad y(1) = 0.$

    Determínense los valores de λ para los cuales existen soluciones no triviales ($y \neq 0$) y escríbanse éstas en función de λ.

Capítulo 8

Métodos numéricos para ecuaciones en derivadas parciales

8.1. Introducción

Dedicamos este capítulo a presentar una breve introducción de algunas de las técnicas que aproximan las soluciones de EDP de segundo orden con coeficientes constantes y dos variables independientes, mostrando cómo pueden aplicarse estas técnicas a ciertos problemas físicos bien conocidos. Estas ecuaciones, como bien sabemos, son de tipo elíptico, parabólico o hiperbólico. El ejemplo típico de ecuación elíptica es la ecuación de Laplace (potencial) para la distribución de la temperatura en estado de equilibrio en una región bidimensional, el de ecuación parabólica es la ecuación del calor, que describe la distribución de temperatura en una varilla fina, y el de ecuación hiperbólica es la ecuación de ondas para una cuerda vibrante. Las técnicas numéricas para resolver estas EDP son principalmente de dos tipos: métodos de diferencias finitas y métodos de elementos finitos.

Comenzaremos describiendo los métodos numéricos basados en la aproximación por diferencias finitas, que son de dos tipos: explícitos e implícitos. En los primeros, las incógnitas se van calculando sucesivamente, en términos de valores conocidos o variables ya calculadas. En los segundos, hay que resolver un sistema de ecuaciones en cada paso para determinar las incógnitas. Los aplicaremos a las ecuaciones elípticas, parabólicas e hiperbólicas utilizando respectivamente la ecuación de Laplace, la ecuación del calor y la ecuación de ondas como ejemplos. Las técnicas aquí desarrolladas se basan en reemplazar las derivadas parciales por aproximaciones en diferencias finitas que conducen a un sistema de ecuaciones algebraicas para los valores de la función incógnita en una colección de puntos.

Terminamos la lección con una breve introducción al método de los elementos finitos (abreviadamente, MEF). Una ventaja de este método sobre los métodos de diferencias finitas es la relativa facilidad con que se manejan las condiciones de contorno del problema. Muchos problemas físicos tienen condiciones de contorno que incluyen derivadas y se formulan sobre regiones cuya frontera es una curva irregular. Condiciones de contorno de este tipo son difíciles de manejar utilizando técnicas de diferencias finitas porque en cada condición de contorno que incluya una derivada, ésta se debe aproximar mediante un cociente incremental en una malla de puntos y la irregularidad de la frontera hace que sea difícil situar los puntos de la malla. En el MEF las condiciones de contorno aparecen como integrales en un funcional que se debe minimizar, de manera que el procedimiento de construcción es independiente de la condición de contorno concreta de cada problema. Ilustraremos este método con la ecuación de Laplace.

8.2. Métodos de diferencias finitas para ecuaciones elípticas

La EDP elíptica que vamos a considerar es la ecuación bidimensional de Laplace:

$$u_{xx} + u_{yy} = 0$$

sobre el dominio rectangular $\mathcal{R} = \{(x, y)/\ a \le x \le b,\ c \le y \le d\}$ y sujeta a las condiciones de contorno

$$u(x, c) = f_1(x), \quad u(x, d) = f_2(x), \quad a \le x \le b,$$

$$u(a, y) = g_1(y), \quad u(b, y) = g_2(y), \quad c \le y \le d.$$

El método que vamos a utilizar es una adaptación del método de diferencias finitas para problemas de contorno lineales tratado en el capítulo 6.

En primer lugar, tomamos dos números naturales n y m y definimos los tamaños de paso h y k mediante $h = \frac{b-a}{n}$ y $k = \frac{d-c}{m}$. Dividimos el intervalo $[a, b]$ en n partes iguales de anchura h y el intervalo $[c, d]$ en m partes iguales de anchura k. Esto nos permite hacer una malla cuadriculada sobre el rectángulo $\mathcal{R}$ trazando líneas verticales y horizontales por los puntos (x_i, y_j), donde

$$x_i = a + ih, \quad y_j = c + jk, \quad i = 0, 1, \ldots, n, \quad j = 0, 1, \ldots, m, \quad \text{(véase la figura 8.1).}$$

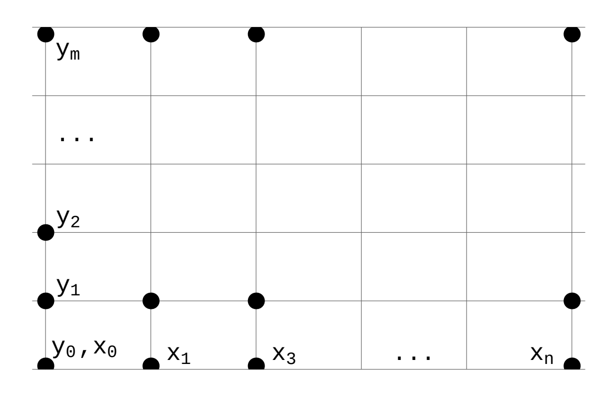


Figura 8.1: Líneas de malla y puntos de malla.

Las líneas rectas $x = x_i$ e $y = y_j$ se llaman **líneas de malla** y sus intersecciones **puntos de malla** de la cuadrícula. Utilizando en cada punto de malla del interior de la cuadrícula la fórmula de Taylor en la variable x alrededor de x_i, generamos la fórmula de diferencias centradas

$$u_{xx}(x_i, y_j) = \frac{u(x_{i-1}, y_j) - 2u(x_i, y_j) + u(x_{i+1}, y_j)}{h^2} + \mathcal{O}(h^2).$$

Con la fórmula de Taylor en la variable y alrededor de y_j generamos la fórmula de diferencias centradas

$$u_{yy}(x_i, y_j) = \frac{u(x_i, y_{j-1}) - 2u(x_i, y_j) + u(x_i, y_{j+1})}{k^2} + \mathcal{O}(k^2).$$

Sustituyendo estas fórmulas en la ecuación de Laplace y eliminando los términos de error $\mathcal{O}(h^2)$ y $\mathcal{O}(k^2)$, obtenemos las siguientes ecuaciones

$$\frac{u(x_{i+1}, y_j) - 2u(x_i, y_j) + u(x_{i-1}, y_j)}{h^2} + \frac{u(x_i, y_{j+1}) - 2u(x_i, y_j) + u(x_i, y_{j-1})}{k^2} = 0,$$

para cada $i = 1, 2, \ldots, n-1$ y $j = 1, 2, \ldots, m-1$. Las condiciones de contorno son

$$u(x_i, y_0) = f_1(x_i), \quad u(x_i, y_m) = f_2(x_i), \quad i = 1, 2, \ldots, n-1,$$

$$u(x_0, y_j) = g_1(y_j), \quad u(x_n, y_j) = g_2(y_j), \quad j = 0, 1, \ldots, m.$$

Si denotamos por $u_{i,j}$ las aproximaciones de los valores $u(x_i, y_j)$, obtenemos la ecuación en diferencias que define el **método de diferencias finitas para la ecuación de Laplace** (véase [7]), cuyo error es de orden $\mathcal{O}(h^2 + k^2)$:

$$2\left[\left(\frac{h}{k}\right)^2 + 1\right] u_{i,j} - (u_{i-1,j} + u_{i+1,j}) - \left(\frac{h}{k}\right)^2 (u_{i,j-1} + u_{i,j+1}) = 0,$$

para cada $i = 1, 2, \ldots, n-1$ y $j = 1, 2, \ldots, m-1$, con

$$u_{i,0} = f_1(x_i), \quad u_{i,m} = f_2(x_i), \quad i = 1, 2, \ldots, n-1,$$
$$u_{0,j} = g_1(y_j), \quad u_{n,j} = g_2(y_j), \quad j = 0, 1, \ldots, m.$$

Una ecuación típica relaciona las aproximaciones de $u(x, y)$ en los puntos

$$(x_{i-1}, y_j), \quad (x_i, y_j), \quad (x_{i+1}, y_j), \quad (x_i, y_{j-1}) \quad \text{y} \quad (x_i, y_{j+1}).$$

Si reporducimos la porción de la malla en la que se localizan estos puntos (véase la figura 8.2), vemos que cada ecuación relaciona las aproximaciones en una región estrellada alrededor de (x_i, y_j).

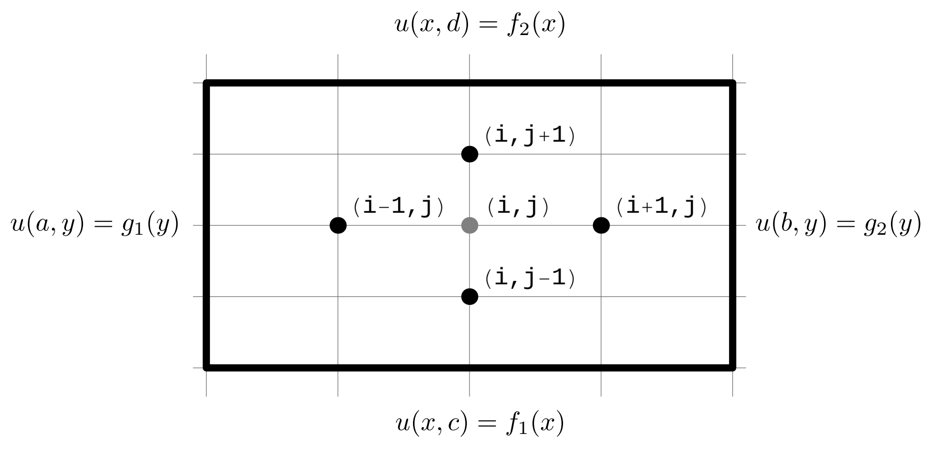


Figura 8.2: Forma esquemática para el método de diferencias finitas (ecuación de Laplace).

Obsérvese que el método anterior proporciona una sistema lineal $(n-1)(m-1) \times (n-1)(m-1)$, cuyas incógnitas son las aproximaciones $u_{i,j}$ de $u(x_i, y_j)$ en los puntos de malla interiores. Este sistema es de gran tamaño y su matriz tiene a lo sumo cinco elementos no nulos en cada fila. Para un uso general, las técnicas iterativas representan a menudo la mejor aproximación a la solución de tales sistemas de ecuaciones. Sin embargo, si el número de ecuaciones no es demasiado grande (del orden de 100 o menor), una solución directa de estos sistemas es práctica, pues el hecho de que la matriz sea definida positiva asegura su estabilidad frente a los errores de redondeo.

Ejemplo. Ilustramos a continuación la solución de la ecuación de Laplace cuando $n = m = 4$. Consideremos el problema de determinar la distribución estacionaria del calor en una fina lámina cuadrada de metal de 2 metros de lado. Dos lados adyacentes se mantienen a 300°C y 200°C, mientras que la temperatura en los otros dos lados se mantiene a 100°C en uno y en el otro crece linealmente de 0°C a 400°C en el vértice donde dichos lados se encuentran. El problema se expresa como

$$u_{xx} + u_{yy} = 0, \quad 0 < x < 2, \quad 0 < y < 2,$$

$$\text{con} \quad \begin{cases} u(x,0) = 300, & u(x,2) = 100, \quad 0 \le x \le 2, \\ u(0,y) = 200, & u(2,y) = 200y, \quad 0 \le y \le 2. \end{cases}$$

Tomando $n = m = 4$ para este problema ($h = k = 0.5$), la malla resultante se muestra en la figura 8.3 y la correspondiente ecuación en diferencias es

$$4u_{i,j} - u_{i+1,j} - u_{i-1,j} - u_{i,j-1} - u_{i,j+1} = 0, \quad i = 1, 2, 3, \quad j = 1, 2, 3.$$

Expresando esto y etiquetando fila por fila, empezando por la esquina inferior izquierda, los valores desconocidos $u_{i,j}$ en los nueve puntos interiores de la malla por $v_1, v_2, \ldots, v_9$, obtenemos que el sistema lineal

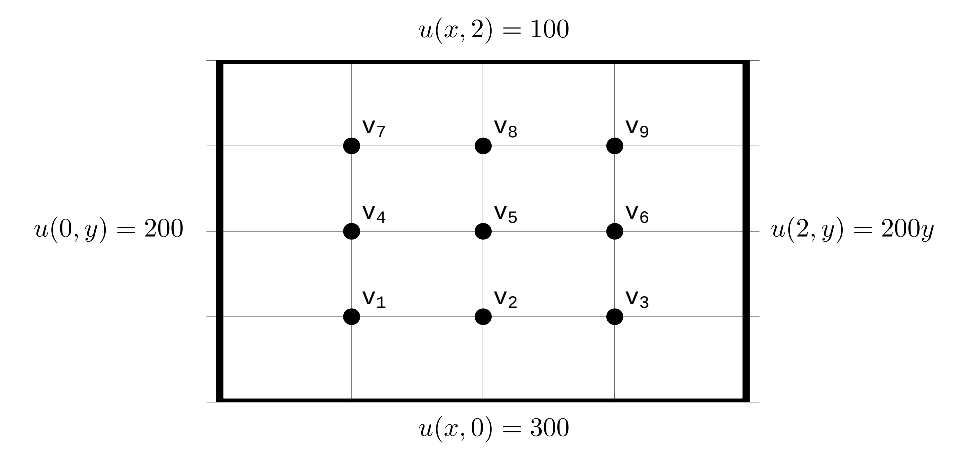


Figura 8.3: Malla de orden 5×5 para una ecuación de Laplace del ejemplo.

asociado al problema es

$$\begin{pmatrix} 4 & -1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 4 & -1 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 4 & 0 & 0 & -1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 4 & -1 & 0 & -1 & 0 & 0 \\ 0 & -1 & 0 & -1 & 4 & -1 & 0 & -1 & 0 \\ 0 & 0 & -1 & 0 & -1 & 4 & 0 & 0 & -1 \\ 0 & 0 & 0 & -1 & 0 & 0 & 4 & -1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & -1 & 4 & -1 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & -1 & 4 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \\ v_6 \\ v_7 \\ v_8 \\ v_9 \end{pmatrix} = \begin{pmatrix} u_{0,1} + u_{1,0} & = & 500 \\ u_{2,0} & = & 300 \\ u_{3,0} + u_{4,1} & = & 400 \\ u_{0,2} & = & 200 \\ & & 0 \\ u_{4,2} & = & 200 \\ u_{0,3} + u_{1,4} & = & 300 \\ u_{2,4} & = & 100 \\ u_{4,3} + u_{3,4} & = & 400 \end{pmatrix},$$

donde el vector de los términos independientes se ha calculado utilizando las condiciones de contorno.

Utilizando el método de eliminación de Gauss, véase [16], la temperatura en los puntos interiores de la malla es

$$v_1 = 233.929, \quad v_2 = 235.714, \quad v_3 = 208.929,$$

$$v_4 = 200.000, \quad v_5 = 200.000, \quad v_6 = 200.000,$$

$$v_7 = 166.071, \quad v_8 = 164.286, \quad v_9 = 191.071.$$

Por supuesto, con un número tan pequeño de puntos de malla no se puede esperar una exactitud alta. Si elegimos un valor de h más pequeño, la exactitud debería mejorar. La figura 8.4 muestra la aproximación numérica anterior de la solución. $\square$

8.3. Métodos de diferencias finitas para ecuaciones parabólicas

Un ejemplo clásico de EDP parabólica es la conocida ecuación del calor, o de la difusión,

$$u_t = \beta u_{xx}, \quad \text{para} \quad 0 < x < \ell \quad \text{y} \quad t > 0,$$

con las condiciones

$$u(0, t) = u(\ell, t) = 0, \quad t > 0, \quad \text{(condiciones de contorno)},$$

$$u(x, 0) = f(x), \quad 0 \leq x \leq \ell, \quad \text{(condición inicial)}.$$

El enfoque que vamos a utilizar para aproximar la solución de este problema emplea diferencias finitas de una manera parecida a como lo hemos hecho en la sección anterior. Empezamos tomando un número natural

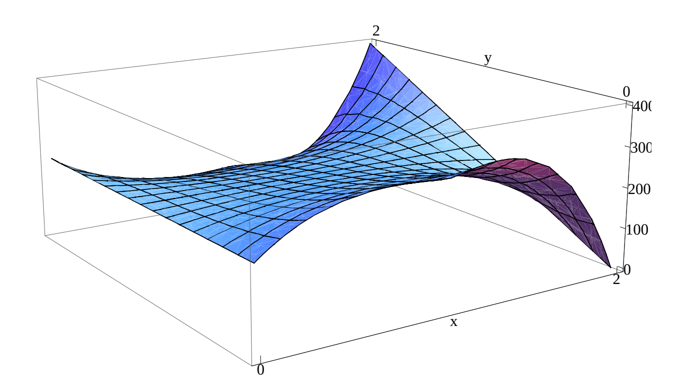


Figura 8.4: Representación gráfica de la solución numérica de la ecuación de Laplace del ejemplo.

$m > 0$ y definiendo $h = \frac{\ell}{m}$. Luego elegimos un tamaño de paso k para la variable temporal t. Los puntos de malla en este caso son (x_i, t_j), donde $x_i = ih$, para $i = 0, 1, \ldots, m$ y $t_j = jk$, para $j = 0, 1, \ldots$

Construimos el método de diferencias utilizando la fórmula de Taylor en t para generar la fórmula de diferencias progresivas

$$u_t(x_i, t_j) = \frac{u(x_i, t_{j+1}) - u(x_i, t_j)}{k} + \mathcal{O}(k),$$

y la fórmula de Taylor en x para generar la fórmula de diferencias centradas

$$u_{xx}(x_i, t_j) = \frac{u(x_{i-1}, t_j) - 2u(x_i, t_j) + u(x_{i+1}, t_j)}{h^2} + \mathcal{O}(h^2).$$

Sustituyendo estas ecuaciones en la EDP, denotando las aproximaciones de los valores $u(x_i, t_j)$ por $u_{i,j}$ y despreciando los términos de error $\mathcal{O}(k)$ y $\mathcal{O}(h^2)$, obtenemos la correspondiente ecuación en diferencias

$$\frac{u_{i,j+1} - u_{i,j}}{k} = \beta \, \frac{u_{i-1,j} - 2u_{i,j} + u_{i+1,j}}{h^2}.$$

Por comodidad, tomamos $\lambda = \frac{\beta k}{h^2}$ en la ecuación en diferencias anterior y reordenamos los términos para obtener la ecuación en diferencias progresiva

$$u_{i,j+1} = (1 - 2\lambda) u_{i,j} + \lambda \left(u_{i-1,j} + u_{i+1,j} \right), \quad i = 1, 2, \ldots, m-1, \quad j = 0, 1, \ldots$$

En forma esquemática, la ecuación anterior puede verse en la figura 8.5. La solución en cada punto $(i, j+1)$ del nivel $(j+1)$-ésimo de tiempo se puede expresar en términos de los valores de la solución en los puntos $(i-1, j)$, (i, j) y $(i+1, j)$ del nivel anterior de tiempo. Fijando un instante final T, elegimos un número de subintervalos temporales n y con la expresión anterior vamos calculando la solución en cada instante hasta llegar a T. Este procedimiento se conoce como **método de diferencias finitas progresivas** o **método explícito clásico**, es un método explícito (ya que todas las aproximaciones pueden hallarse directamente a partir de la información dada por las condiciones iniciales y las de contorno) y de orden $\mathcal{O}(k + h^2)$.

Los valores de la condición inicial $u(x_i, 0) = f(x_i)$, para $i = 0, 1, \ldots, m$, se utilizan en la ecuación en diferencias para hallar los valores de $u_{i,1}$, para $i = 1, 2, \ldots, m-1$. Las condiciones de contorno $u(0, t) = u(\ell, t) = 0$, implican que $u_{0,1} = u_{m,1} = 0$, así que podemos determinar todas las aproximaciones de la forma $u_{i,1}$. Aplicando el mismo procedimiento, una vez que se conocen todas las aproximaciones $u_{i,1}$, podemos calcular los valores $u_{i,2}, u_{i,3}, \ldots, u_{i,m-1}$ de forma parecida.

La naturaleza explícita del método implica que la matriz $(m-1) \times (m-1)$ asociada es tridiagonal:

$$\begin{pmatrix} 1-2\lambda & \lambda & & & & \\ \lambda & 1-2\lambda & \lambda & & & \\ & \lambda & 1-2\lambda & \lambda & & \\ & & \ddots & \ddots & \ddots & \\ & & & \lambda & 1-2\lambda & \lambda \\ & & & & \lambda & 1-2\lambda \end{pmatrix}.$$

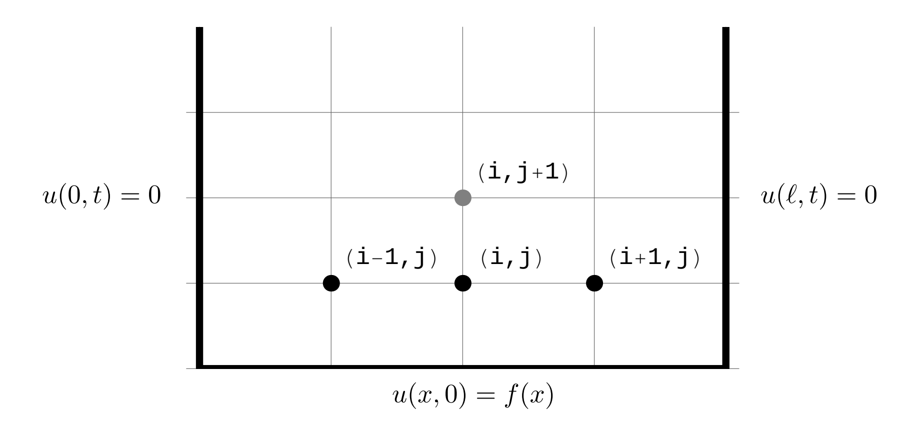


Figura 8.5: Forma esquemática para el método explícito clásico (ecuación del calor).

Si definimos $\mathbf{u}^{(0)} = (f(x_1), f(x_2), \dots, f(x_{m-1}))^T$ y

$$\mathbf{u}^{(j)} = (u_{1,j}, u_{2,j}, \dots, u_{m-1,j})^T, \quad \text{para cada } j = 1, 2, \dots$$

entonces la solución aproximada viene dada por

$$\mathbf{u}^{(j)} = A\mathbf{u}^{(j-1)}, \quad \text{para cada } j = 1, 2, \dots$$

de manera que $\mathbf{u}^{(j)}$ se obtiene a partir de $\mathbf{u}^{(j-1)}$ mediante una simple multiplicación matricial

Comentario adicional. ▷ El método explícito anterior no necesariamente produce buenos resultados. Puede verse en [16] que el método no refleja la exactitud esperada de orden $\mathcal{O}(k+h^2)$ para un problema concreto. Esto se debe a la condición de estabilidad del método explícito. Hay que hacer elecciones apropiadas para los tamaños de paso h y k que determinan los valores de λ. Se puede demostrar, véase [15], que el método explícito es estable si λ es tal que $0 < \lambda \leq \frac{1}{2}$. Esto significa que el tamaño de paso k debe cumplir que $k \leq \frac{h^2}{2\beta}$, de manera que si esto no se cumple, puede ocurrir que los errores introducidos en una fila j, $\{u_{ij}\}$, se incrementen en alguna fila posterior $\{u_{ir}\}$ para algún $r > j$.

Para obtener un método más estable, consideraremos un método **implícito**. La ecuación en diferencias finitas de este método se obtiene al discretizar la EDP en dos instantes y hacer la media. Así, si en la EDP reemplazamos $u_t(x,t)$ por una diferencia progresiva y $u_{xx}(x,t)$ por el valor medio de una diferencia centrada en los pasos de tiempo j y $j+1$, da

$$\frac{u_{i,j+1} - u_{i,j}}{k} = \frac{\beta}{2}\left[\frac{u_{i-1,j} - 2u_{i,j} + u_{i+1,j}}{h^2} + \frac{u_{i-1,j+1} - 2u_{i,j+1} + u_{i+1,j+1}}{h^2}\right],$$

que tiene un error de orden $\mathcal{O}(k^2 + h^2)$. Tomando de nuevo $\lambda = \frac{\beta k}{h^2}$, la ecuación en diferencias anterior se puede escribir ahora como

$$-\lambda u_{i-1,j+1} + 2(1+\lambda)u_{i,j+1} - \lambda u_{i+1,j+1} = \lambda u_{i-1,j} + 2(1-\lambda)u_{i,j} + \lambda u_{i+1,j},$$

para $i = 1, 2, \dots, m-1$ y $j = 0, 1, \dots$ Este procedimiento se llama **método de Crank-Nicolson** y es incondicionalmente estable. Su forma esquemática puede verse en la figura 8.6.

La forma matricial del método de Crank-Nicolson es

$$A\mathbf{u}^{(j+1)} = B\mathbf{u}^{(j)}, \quad j = 0, 1, 2, \dots$$

donde $\mathbf{u}^{(j)} = (u_{1,j}, u_{2,j}, \dots, u_{m-1,j})^T$,

$$A = \begin{pmatrix} 2(1+\lambda) & -\lambda & & & & \\ -\lambda & 2(1+\lambda) & -\lambda & & & \\ & -\lambda & 2(1+\lambda) & -\lambda & & \\ & & \ddots & \ddots & \ddots & \\ & & & -\lambda & 2(1+\lambda) & -\lambda \\ & & & & -\lambda & 2(1+\lambda) \end{pmatrix},$$

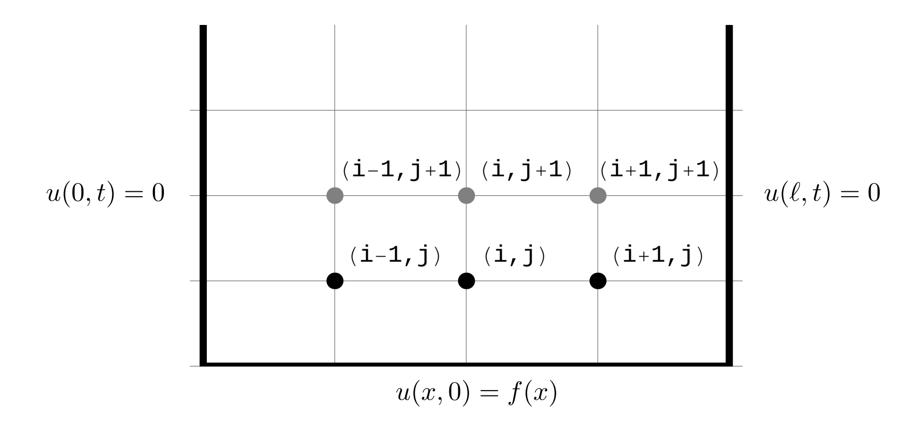


Figura 8.6: Forma esquemática para el método de Crank-Nicolson (ecuación del calor).

$$B=\begin{pmatrix} 2(1-\lambda) & \lambda & & & & \\ \lambda & 2(1-\lambda) & \lambda & & & \\ & \lambda & 2(1-\lambda) & \lambda & & \\ & & \ddots & \ddots & \ddots & \\ & & & \lambda & 2(1-\lambda) & \lambda \\ & & & & \lambda & 2(1-\lambda) \end{pmatrix}.$$

Obsérvese que el término del miembro derecho de la ecuación matrical anterior es conocido, así que esta ecuación es un sistema lineal tridiagonal. La matriz tridiagonal A es definida positiva y diagonal estrictamente dominante, así que es no singular y el sistema de ecuaciones se puede entonces resolver mediante cualquier método descrito en el capítulo 1.

Ejemplo. Consideramos la ecuación del calor

$$u_t = u_{xx} = 0, \quad 0 < x < 1, \quad t > 0,$$

$$\text{con} \quad \begin{cases} u(0,t) = u(1,t) = 0, \quad t > 0, \quad \text{(condiciones de contorno)}, \\ u(x,0) = \operatorname{sen} \pi x, \quad 0 \le x \le 1, \quad \text{(condición inicial)}. \end{cases}$$

Vamos a tomar $h = 0.2$ y $k = 0.05$, de manera que $\lambda = 1.25$ y $m = 5$. Reemplazando el valor de λ en la ecuación en diferencias, obtenemos

$$-1.25u_{i-1,j+1} + 4.5u_{i,j+1} - 1.25u_{i+1,j+1} = 1.25u_{i-1,j} - 0.5u_{i,j} + 1.25u_{i+1,j}, \quad i = 1,2,3,4.$$

En el primer paso de tiempo $t = k$, $u_{i,1}$ está dado por la solución del sistema tridiagonal

$$\begin{pmatrix} 4.5 & -1.25 & 0 & 0 \\ -1.25 & 4.5 & -1.25 & 0 \\ 0 & -1.25 & 4.5 & -1.25 \\ 0 & 0 & -1.25 & 4.5 \end{pmatrix} \begin{pmatrix} u_{1,1} \\ u_{2,1} \\ u_{3,1} \\ u_{4,1} \end{pmatrix} = \begin{pmatrix} -0.5u_{1,0} + 1.25u_{2,0} \\ 1.25u_{1,0} - 0.5u_{2,0} + 1.25u_{3,0} \\ 1.25u_{2,0} - 0.5u_{3,0} + 1.25u_{4,0} \\ 1.25u_{3,0} - 0.5u_{4,0} \end{pmatrix}$$

$$= \begin{pmatrix} 0.894928 \\ 1.448024 \\ 1.448024 \\ 0.894928 \end{pmatrix},$$

donde $u_{i,0} = \operatorname{sen}(\pi i h)$. La solución del sistema tridiagonal es

$$u(x_i, 0.05) = (0.36122840, 0.58447983, 0.58447983, 0.36122840)^T, \quad i = 1,2,3,4.$$

La solución aproximada en $t = 0.5$, después de 10 pasos de tiempo, puede encontrarse en [16]. La figura 8.7 muestra la aproximación numérica de la solución. $\square$

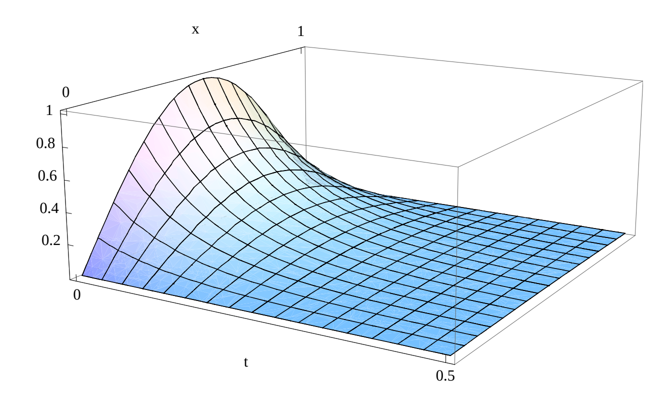


Figura 8.7: Representación gráfica de la solución numérica de la ecuación del calor del ejemplo.

8.4. Métodos de diferencias finitas para ecuaciones hiperbólicas

Como ejemplo de una EDP hiperbólica, consideramos la ecuación de ondas

$$u_{tt} = \alpha^2 \, u_{xx}, \quad 0 < x < \ell, \quad t > 0,$$

sujeta a las condiciones

$$u(0,t) = u(\ell,t) = 0, \quad t > 0, \quad \text{(condiciones de contorno)},$$

$$u(x,0) = f(x), \quad u_t(x,0) = g(x), \quad 0 \leq x \leq \ell, \quad \text{(condiciones iniciales)}.$$

Tomamos un número natural $m > 0$ y un tamaño de paso para la variable temporal $k > 0$. Con $h = \frac{\ell}{m}$, los puntos de malla (x_i, t_j) están dados por $x_i = ih$ y $t_j = jk$, para cada $i = 0, 1, \ldots, m$ y $j = 0, 1, \ldots$ El método de diferencias finitas se obtiene utilizando las fórmulas de diferencias centradas para las derivadas parciales segundas dadas por

$$u_{tt}(x_i, t_j) = \frac{u(x_i, t_{j-1}) - 2u(x_i, t_j) + u(x_i, t_{j+1})}{k^2} + \mathcal{O}(k^2),$$

$$u_{xx}(x_i, t_j) = \frac{u(x_{i-1}, t_j) - 2u(x_i, t_j) + u(x_{i+1}, t_j)}{h^2} + \mathcal{O}(h^2).$$

Sustituyendo estas fórmulas en la ecuación de ondas, denotando las aproximaciones de los valores $u(x_i, t_j)$ por $u_{i,j}$y despreciando los términos de error $\mathcal{O}(k^2)$ y $\mathcal{O}(h^2)$, obtenemos la ecuación en diferencias de orden $\mathcal{O}(k^2 + h^2)$

$$\frac{u_{i,j-1} - 2u_{i,j} + u_{i,j+1}}{k^2} = \alpha^2 \, \frac{u_{i-1,j} - 2u_{i,j} + u_{i+1,j}}{h^2}.$$

Tomando $\lambda = \alpha k/h$, podemos despejar $u_{i,j+1}$, la aproximación más avanzada en el tiempo, para obtener la fórmula **explícita**

$$u_{i,j+1} = 2(1 - \lambda^2)u_{i,j} + \lambda^2(u_{i-1,j} + u_{i+1,j}) - u_{i,j-1}, \quad i = 1, 2, \ldots, m-1, \quad j = 1, 2, \ldots$$

Esta fórmula se muestra de forma esquemática en la figura 8.8. La solución en cada punto $(i, j+1)$ del nivel $(j+1)$-ésimo de tiempo está expresada en términos de los valores solución en los puntos $(i-1, j)$, (i, j), $(i+1, j)$ y $(i, j-1)$ de los dos niveles de tiempo precedentes. Dicha fórmula tiene problemas de estabilidad y se puede demostrar, véase [15], que el método es estable si $0 < \lambda \leq 1$.

Observemos que la expresión anterior nos permite obtener la soluciónen el instante t_{j+1} a partir de la solución en los instantes t_j y t_{j-1}. Es decir, para calcular la entrada $u_{i,j+1}$ en el nivel de tiempo $(j+1)$, debemos conocer las entradas de los niveles de tiempo j y $(j-1)$. Esto supone un pequeño problema de partida porque solo conocemos la primera fila de la condición inicial $u_{i,0} = f(x_i)$. Para obtener la segunda

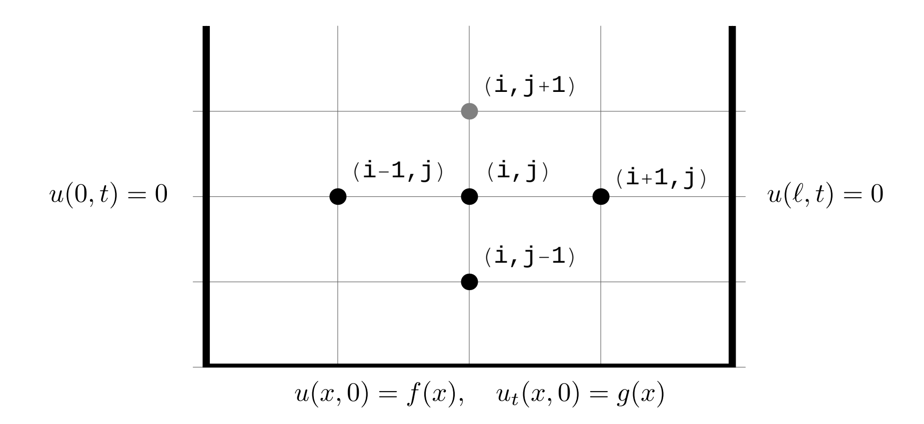


Figura 8.8: Forma esquemática para el método de diferencias finitas con tres niveles (ecuación de ondas).

fila corresponiente a $u_{i,1}$, hay que utilizar la segunda condición inicial $u_t(x,0) = g(x)$. Una posibilidad es sustituir u_t por una aproximación en diferencias progresivas

$$u_t(x_i,0) = \frac{u(x_i,t_1) - u(x_i,0)}{k} + \mathcal{O}(k),$$

que nos permite obtener una ecuación en diferencias finitas que da una aproximación para la segunda fila con un error de truncamiento de solo $\mathcal{O}(k)$. Para obtener una mejor aproximación, consideramos el desarrollo en serie de Taylor de $u(x,t)$ alrededor del punto $(x_i,0)$

$$u(x_i,t_1) = u(x_i,k) = u(x_i,0) + ku_t(x_i,0) + \frac{k^2}{2}u_{tt}(x_i,0) + \mathcal{O}(k^3).$$

Suponiendo que la derivada segunda de $f(x)$ existe, tenemos

$$u_{tt}(x_i,0) = \alpha^2 u_{xx}(x_i,0) = \alpha^2 f''(x_i),$$

de manera que, utilizando las condiciones iniciales $u_t(x_i,0) = g(x_i)$ y $u(x_i,0) = f(x_i)$, se sigue que

$$u(x_i,t_1) = f(x_i) + kg(x_i) + \frac{k^2}{2}\alpha^2 f''(x_i) + \mathcal{O}(k^3).$$

Si no es posible calcular $f''(x_i)$ directamente, podemos reemplazar $f''(x_i)$ en la última ecuación por una fórmula de diferencias centradas

$$f''(x_i) = \frac{f(x_{i-1}) - 2f(x_i) + f(x_{i+1})}{h^2} + \mathcal{O}(h^2).$$

En este caso, la aproximación numérica en la segunda fila está dada por la fórmula

$$\begin{aligned} u_{i,1} &= f(x_i) + kg(x_i) + \frac{k^2\alpha^2}{2h^2}(f(x_{i-1}) - 2f(x_i) + f(x_{i+1})) \\ &= (1-\lambda^2)f(x_i) + \frac{\lambda^2}{2}(f(x_{i-1}) + f(x_{i+1})) + kg(x_i), \quad i = 1,2,\ldots,m-1, \end{aligned}$$

que tiene una exactitud de orden $\mathcal{O}(k^3 + h^2k^2)$.

Ejemplo. Sea la ecuación de ondas

$$u_{tt} = 16\, u_{xx}, \quad 0 < x < 1, \quad t > 0,$$

sujeta a las condiciones

$$u(0,t) = u(1,t) = 0, \quad t > 0, \quad \text{(condiciones de contorno)},$$

$$u(x,0) = \operatorname{sen} \pi x, \quad u_t(x,0) = 0, \quad 0 \le x \le 1, \quad \text{(condiciones iniciales)}.$$

Tomamos $h = 0.2$ y $k = 0.05$, de manera que $\lambda = 1$ y $m = 5$. Las aproximaciones de u en $t = 0.05$, para $i = 1, 2, 3, 4$, son como se sigue a continuación. Las condiciones de contorno dan

$$u_{0,j} = u_{5,j} = 0, \quad j = 1, 2, \ldots$$

y las condiciones iniciales dan

$$\begin{aligned} u_{i,0} &= \operatorname{sen}(0.2\pi i), \quad i = 0, 1, 2, 3, 4, 5, \\ u_{i,1} &= \tfrac{1}{2}(f(0.2(i-1)) + f(0.2(i+1))) + (0.05)g(0.2i) \\ &= 0.5[\operatorname{sen}(0.2\pi(i-1)) + \operatorname{sen}(0.2\pi(i+1))], \quad i = 1, 2, 3, 4. \end{aligned}$$

Luego,

$$u_{1,1} = 0.47552826, \quad u_{2,1} = 0.76942088, \quad u_{3,1} = 0.76942088, \quad u_{4,1} = 0.47552826.$$

Para $t = 2k = 0.1$, obtenemos la ecuación en diferencias

$$u_{i,2} = u_{i-1,1} + u_{i+1,1} - u_{i,0}, \quad i = 1, 2, 3, 4,$$

que implica que,

$$\begin{aligned} u_{1,2} &= u_{2,1} - u_{1,0} = 0.18163563, \\ u_{2,2} &= u_{1,1} + u_{3,1} - u_{2,0} = 0.29389263, \\ u_{3,2} &= u_{2,1} + u_{4,1} - u_{3,0} = 0.29389263, \\ u_{4,2} &= u_{3,1} - u_{4,0} = 0.18163563. \end{aligned}$$

La solución aproximada en $t = 0.5$, después de 10 pasos de tiempo, junto con una representación en tres dimensiones, se puede encontrar en [16]. La figura 8.9 muestra la aproximación numérica de la solución. □

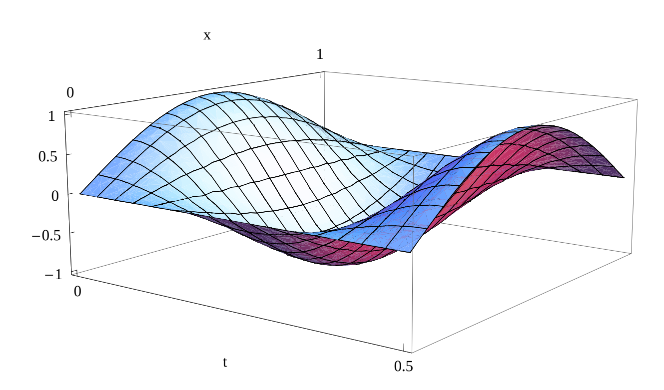


Figura 8.9: Representación gráfica de la solución numérica de la ecuación de ondas del ejemplo.

8.5. Introducción al método de los elementos finitos

En las secciones anteriores se han descrito métodos de diferencias finitas para resolver problemas de contorno con EDP. Para estos métodos los puntos de la malla se suelen colocar uniformemente espaciados dentro de la región. Si los puntos no están exactamente en las fronteras, se requieren ajustes especiales para los puntos adyacentes a la frontera y los errores son mayores cuando se hace esto.

En esta sección vamos a introducir un método diferente especialmente valioso para regiones irregulares: el método de los elementos finitos (abreviadamente, MEF). En este método los puntos pueden estar espaciados de manera no uniforme. Esto no solo resuelve el problema de hacer coincidir los puntos con una frontera, sino que también facilita la colocación más estrecha de los puntos entre sí en partes de la región donde la solución del problema varía rápidamente, con lo que se mejora la exactitud. Hoy día matemáticos e ingenieros utilizan ampliamente el MEF.

Introducimos a continuación el MEF para EDP elípticas. La EDP elíptica que vamos a resolver en esta sección es la ecuación bidimensional de Laplace,

$$u_{xx} + u_{yy} = 0,$$

con $(x, y) \in \mathcal{S}$, donde $\mathcal{S}$ es una región plana.

Un ejemplo típico es la determinación de la distribución del potencial eléctrico $u(x, y)$ en el espacio entre dos conductores rectangulares. Dada la simetría del problema, solo una cuarta parte de la región real del problema necesita analizarse. En este caso, surgen dos clases de condiciones en la frontera (véase la figura 8.10): valores de potencial constante a lo largo de las superficies conductoras (llamadas *condiciones de Dirichlet*) y valores nulos de derivadas normales a lo largo de los planos de simetría. Es decir, se tienen las condiciones en los conductores

$$u(x, y) = \begin{cases} 0 & \text{en el conductor interior,} \\ 1 & \text{en el conductor exterior,} \end{cases}$$

y en las superficies normales a los circuitos

$$\frac{\partial u}{\partial n} = 0.$$

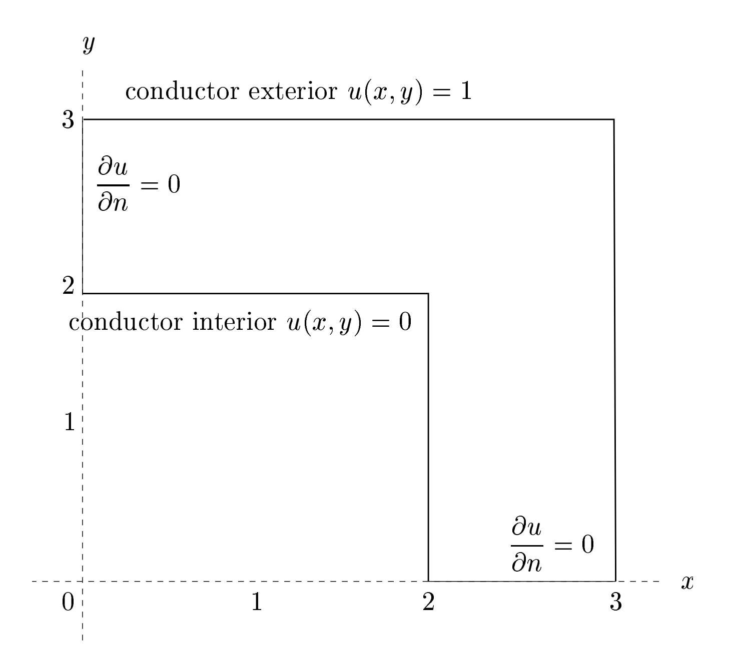


Figura 8.10: Región plana $\mathcal{S}$ con condiciones de contorno.

El *principio de la energía de potencial mínima* dice que $u(x, y)$ es solución si y solo si minimiza la función energía:

$$I[u] = \frac{1}{2} \int \int_{\mathcal{S}} \left[\left(\frac{\partial u}{\partial x} \right)^2 + \left(\frac{\partial u}{\partial y} \right)^2 \right] dxdy,$$

donde $\mathcal{S}$ es la región plana de la figura 8.10.

El MEF minimiza la función energía $I[u]$ asociada al potencial $u(x, y)$ suponiendo que el pontencial $u(x, y)$ está dado por una combinación lineal de funciones elementales (usualmente polinomios de grado fijo en las variables x e y) con coeficientes a determinar y definidas a trozos sobre una partición de $\mathcal{S}$ que consideremos.

En nuestro caso, utilizaremos los denominados *elementos triangulares de primer orden*. Para ello, dividimos la región $\mathcal{S}$ en triángulos T_i como en la figura 8.11 (cuantos más triángulos, mejor será la aproximación) y denotamos los vértices por E_i.

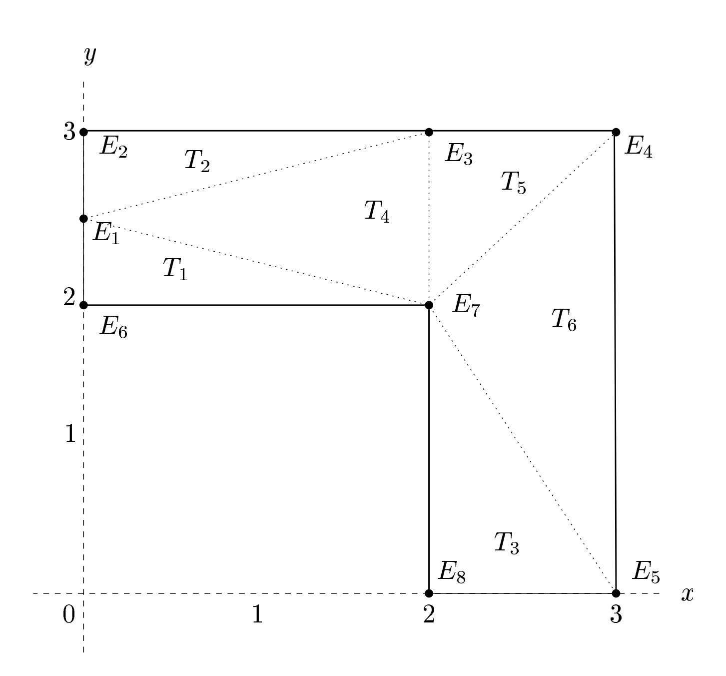


Figura 8.11: Región plana $\mathcal{S}$ dividida en triángulos.

El MEF busca una aproximación de la forma

$$U(x,y) = \sum_{i=1}^{8} \gamma_i \, u_i(x,y),$$

donde γ_i son constantes, el número de sumandos coincide con el número de vértices E_i y $u_i(x,y)$ es tal que

1. sobre cada triángulo T_i, son polinomios lineales en x e y de la forma $a + bx + cy$, donde a, b y c son constantes distintas por lo general para cada triángulo,
2. $u_i(E_i) = 1$,
3. $u_i(E_j) = 0$ para $j \neq i$.

Si imponemos las condiciones en la frontera, que conocemos para la solución $u(x,y)$, para la aproximación $U(x,y)$, se tiene que

$$U(E_2) = U(E_3) = U(E_4) = U(E_5) = 1 \qquad \text{y} \qquad U(E_6) = U(E_7) = U(E_8) = 0.$$

Entonces, como $u_i(E_k) = 0$ si $k \neq i$ y $u_i(E_i) = 1$, se sigue que

$$\gamma_2 = \gamma_3 = \gamma_4 = \gamma_5 = 1 \qquad \text{y} \qquad \gamma_6 = \gamma_7 = \gamma_8 = 0,$$

por lo que

$$U(x,y) = \gamma_1 u_1(x,y) + \sum_{i=2}^{5} u_i(x,y).$$

A continuación, se calculan las funciones u_i. Calculamos como ejemplo la función u_1. Sobre cada triángulo T_i la función u_1 es de la forma

$$u_1(x,y) = a_i + b_i x + c_i y, \quad \text{para } (x,y) \in T_i.$$

Hay dos clases de triángulos: los que contienen a E_1, que son T_1 y T_2, y los que no lo contienen, que son el resto.

En T_1 los vértices son $E_1 = (0, 2.5), E_6 = (0, 2)$ y $E_7 = (2, 2)$. Así,

$$\begin{aligned} u_1(E_1) &= a_1 + b_1\, 0 + c_1\, 2.5 &= 1, \\ u_1(E_6) &= a_1 + b_1\, 0 + c_1\, 2 &= 0, \\ u_1(E_7) &= a_1 + b_1\, 2 + c_1\, 2 &= 0, \end{aligned}$$

cuya solución es

$$a_1 = -4, \qquad b_1 = 0, \qquad c_1 = 2.$$

Luego,

$$u_1(x, y) = -4 + 2y, \quad \text{para } (x, y) \in T_1.$$

En T_2 los vértices son $E_1 = (0, 2.5)$, $E_2 = (0, 3)$ y $E_3 = (2, 3)$, de manera que

$$\begin{aligned} u_1(E_1) &= a_2 + b_2\, 0 + c_2\, 2.5 &= 1, \\ u_1(E_2) &= a_2 + b_2\, 0 + c_2\, 3 &= 0, \\ u_1(E_3) &= a_2 + b_2\, 2 + c_2\, 3 &= 0. \end{aligned}$$

La solución es

$$a_2 = 6, \qquad b_2 = 0, \qquad c_2 = -2,$$

por lo que

$$u_1(x, y) = 6 - 2y, \quad \text{para } (x, y) \in T_2.$$

En los triángulos T_i que no contienen a E_1 los correspondientes sistemas de ecuaciones que resultan son homogéneos (todos los términos independientes son 0) y libres, de forma que la solución es

$$a_i = 0, \qquad b_i = 0, \qquad c_i = 0$$

y, en consecuencia,

$$u_1(x, y) = 0, \quad \text{para } (x, y) \in T_i \text{ y } E_1 \notin T_i.$$

Resumiendo,

$$u_1(x, y) = \begin{cases} -4 + 2y & \text{para } (x, y) \in T_1 \\ 6 - 2y & \text{para } (x, y) \in T_2 \\ 0 & \text{para } (x, y) \in T_i \text{ y } i \neq 1, 2. \end{cases} \tag{8.1}$$

Se deja como ejercicio para los estudiantes el cálculo de los restantes u_i.

Una vez encontrados todos los u_i solo quedan por determinar los coeficientes γ_i que no se pudieron encontrar al imponer las condiciones de contorno. En nuestro caso, únicamente γ_1. Para ello, imponemos que se minimice la función energía

$$\begin{aligned} I[U] &= \frac{1}{2} \int\int_{\mathcal{S}} \left[\left(\frac{\partial U}{\partial x} \right)^2 + \left(\frac{\partial U}{\partial y} \right)^2 \right] dxdy \\ &= \frac{1}{2} \int\int_{\mathcal{S}} \left[\left(\gamma_1 \frac{\partial u_1}{\partial x} + \sum_{i=2}^{i=5} \frac{\partial u_i}{\partial x} \right)^2 + \left(\gamma_1 \frac{\partial u_1}{\partial y} + \sum_{i=2}^{i=5} \frac{\partial u_i}{\partial y} \right)^2 \right] dxdy. \end{aligned}$$

Para encontrar el mínimo de $I[U]$ es necesario resolver

$$\frac{\partial I}{\partial \gamma_1} = 0.$$

Al derivar, obtenemos

$$\begin{aligned} \frac{\partial I}{\partial \gamma_1} &= \int\int_{\mathcal{S}} \left[\frac{\partial u_1}{\partial x} \left(\gamma_1 \frac{\partial u_1}{\partial x} + \sum_{i=2}^{i=5} \frac{\partial u_i}{\partial x} \right) + \frac{\partial u_1}{\partial y} \left(\gamma_1 \frac{\partial u_1}{\partial y} + \sum_{i=2}^{i=5} \frac{\partial u_i}{\partial y} \right) \right] dxdy \\ &= \gamma_1 \int\int_{\mathcal{S}} \left[\left(\frac{\partial u_1}{\partial x} \frac{\partial u_1}{\partial x} + \frac{\partial u_1}{\partial y} \frac{\partial u_1}{\partial y} \right) dxdy + \sum_{i=2}^{i=5} \int\int_{\mathcal{S}} \left(\frac{\partial u_i}{\partial x} \frac{\partial u_1}{\partial x} + \frac{\partial u_i}{\partial y} \frac{\partial u_1}{\partial y} \right) \right] dxdy, \end{aligned}$$

e igualando a cero, vemos que

$$\gamma_1 = \frac{-\sum_{i=2}^{i=5}\int\int_{\mathcal{S}}\left(\frac{\partial u_i}{\partial x}\frac{\partial u_1}{\partial x}+\frac{\partial u_i}{\partial y}\frac{\partial u_1}{\partial y}\right)dxdy}{\int\int_{\mathcal{S}}\left(\frac{\partial u_1}{\partial x}\frac{\partial u_1}{\partial x}+\frac{\partial u_1}{\partial y}\frac{\partial u_1}{\partial y}\right)dxdy}.$$

Es necesario entonces el cálculo de la suma de cuatro integrales y la integral del denominador.

Como ejemplo calculamos la integral del denominador. Para las integrales del numerador se procede de igual forma, necesitando para ello el cálculo de los u_i con $i > 2$.

Como la región $\mathcal{S}$ se encuentra dividida en triángulos, obtenemos que

$$\begin{aligned}\int\int_{\mathcal{S}}\left(\frac{\partial u_1}{\partial x}\frac{\partial u_1}{\partial x}+\frac{\partial u_1}{\partial y}\frac{\partial u_1}{\partial y}\right)dxdy &= \sum_{i=1}^{6}\int\int_{T_i}\left(\frac{\partial u_1}{\partial x}\frac{\partial u_1}{\partial x}+\frac{\partial u_1}{\partial y}\frac{\partial u_1}{\partial y}\right)dxdy\\ &= \sum_{i=1}^{2}\int\int_{T_i}\left(\frac{\partial u_1}{\partial x}\frac{\partial u_1}{\partial x}+\frac{\partial u_1}{\partial y}\frac{\partial u_1}{\partial y}\right)dxdy,\end{aligned}$$

donde la segunda igualdad se sigue de que la función u_1 es nula sobre los triángulos T_3, T_4, T_5 y T_6 (véase (8.1)).

De nuevo por (8.1) en el triángulo T_1 se tiene que

$$\frac{\partial u_1}{\partial x} = 0, \quad \frac{\partial u_1}{\partial y} = 2,$$

y en el triángulo T_2

$$\frac{\partial u_1}{\partial x} = 0, \quad \frac{\partial u_1}{\partial y} = -2.$$

Por tanto, la integral queda

$$\int\int_{T_1}(0+2\,.\,2)dxdy + \int\int_{T_2}(0+(-2)(-2)dxdy = 4\int\int_{T_1}dxdy + 4\int\int_{T_2}dxdy.$$

Obsérvese que las dos últimas integrales son las áreas de T_1 y T_2 respectivamente, que son igual a $1/2$ (véase la figura 8.11). En consecuencia,

$$\int\int_{\mathcal{S}}\left(\frac{\partial u_1}{\partial x}\frac{\partial u_1}{\partial x}+\frac{\partial u_1}{\partial y}\frac{\partial u_1}{\partial y}\right)dxdy = 4.$$

Terminamos dejando para los estudiantes dos ejercicios. El primero, el cálculo de las integrales

$$\int\int_{\mathcal{S}}\left(\frac{\partial u_i}{\partial x}\frac{\partial u_1}{\partial x}+\frac{\partial u_i}{\partial y}\frac{\partial u_1}{\partial y}\right)dxdy, \quad i = 2,3,4,5.$$

Y el segundo, el cálculo del valor γ_1 que minimiza la función energía y la solución aproximada $U(x, y)$.

8.6. Sugerencias para seguir leyendo

El tratamiento numérico dado en este capítulo a la resolución de EDP es solo una muy breve introducción a un área muy extensa de investigaciones y aplicaciones. A continuación ofrecemos solo algunos de los textos que proporcionan una fuente excelente para un estudio posterior de estos tópicos: Golub y Ortega (1992), Larsson y Thomée (2003), Grossmann y otros (2007) y Gockenbach (2011) y Davis (2011).

8.7. Ejercicios

1. Determínese el sistema de cuatro ecuaciones con cuatro incógnitas v_1, v_2, v_3 y v_4 que se usa para calcular las aproximaciones a la solucón de la ecuación bidimensional de Laplace en el cuadrado $D = \{(x, y)/\ 0 \leq x \leq 3,\ 0 \leq y \leq 3\}$. Los valores en la frontera son

$$u(x,0) = 10, \quad u(x,3) = 90, \quad 0 \leq x \leq 3,$$
$$u(0,y) = 70, \quad u(3,y) = 0, \quad 0 \leq y \leq 3.$$

2. Calcúlense las tres primeras filas de la malla que se construye para la ecuación de Poisson

$$\begin{cases} u_{xx} + u_{yy} = 6x, \quad 0 \le x \le 1, \quad 0 \le y \le 1, \\ u(x,0) = x^3, \quad u(x,1) = x^3 - 3x + 1, \quad 0 \le x \le 1, \\ u(0,y) = y, \quad u(1,y) = -2y + 1, \quad 0 \le y \le 1, \end{cases}$$

utilizando una malla con $h = k = 0.1$. Realícense las operaciones a mano (o con calculadora).

3. Utilícese el método explícito clásico para calcular las tres primeras filas de la malla que se construye para la siguiente ecuación del calor

$$\begin{cases} u_t = u_{xx}, \quad 0 < x < 1, \quad t > 0, \\ u(0,t) = u(1,t) = 0, \quad t > 0, \\ u(x,0) = \operatorname{sen} \pi x, \quad 0 \le x \le 1. \end{cases}$$

Tómese $h = 0.1$ y $k = 0.01$. Realícense las operaciones a mano (o con calculadora).

4. Obsérvese en el método explícito clásico para la ecuación del calor que si la elección de $h = k = \frac{1}{2}$ da un valor de λ mayor que $\frac{1}{2}$, tendríamos que aumentar h o disminuir k. Además, si elegiésemos $\lambda = \frac{1}{2}$ y luego hiciésemos una elección conveniente de k, podría obtenerse una elección irracional de h que no produjese un número entero de subintervalos. Entonces, si $\beta = 3$, $\ell = 4.2$ y se obtiene $\lambda = \frac{1}{4}$, ¿cuáles deberían ser las elecciones de h y k si queremos aproximar $u(x,3)$?

5. Sea $k = \frac{h^2}{2\beta}$.

   a) Utilícese la igualdad anterior en la fórmula de la ecuación en diferencias del método de Crank-Nicolson y simplifíquese la ecuación resultante.

   b) Exprésense las ecuaciones del apartado anterior matricialmente.

   c) ¿Es la matriz correspondiente al apartado anterior diagonal estrictamente dominante?

6. Sea la ecuación parabólica $u_t - u_{xx} = h(x)$.

   a) Desarróllese el método de diferencias finitas progresivas para la ecuación anterior.

   b) Desarróllese el método implícito clásico para la ecuación anterior.

   c) Desarróllese el método de Crank-Nicolson para la ecuación anterior.

7. Si en la ecuación del calor reemplazamos $u_t(x,t)$ por una diferencia regresiva y $u_{xx}(x,t)$ por una diferencia centrada en el j-ésimo paso de tiempo, obtenemos el *método implícito clásico* para la ecuación del calor. Demuéstrese que dicho método está dado por

$$u_{i,j-1} = -\lambda u_{i-1,j} + (1 + 2\lambda) u_{i,j} - \lambda u_{i+1,j}, \quad i = 1, 2, \ldots, m-1, \quad j = 1, 2, \ldots,$$

donde $\lambda = \frac{\beta k}{h^2}$. (Este método es estable y el error es de orden $\mathcal{O}(k + h^2)$.)

8. Utilícese el método implícito clásico obtenido en el ejercicio anterior para aproximar la solución del problema

$$\begin{cases} u_t = \frac{1}{16} u_{xx}, \quad 0 < x < 1, \quad t > 0, \\ u(0,t) = u(1,t) = 0, \quad t > 0, \\ u(x,0) = 2 \operatorname{sen}(2\pi x), \quad 0 \le x \le 1, \end{cases}$$

con $m = 3$ y $k = 0.01$. Calcúlense las tres primeras filas de la malla y compárense los resultados con la solución exacta $u(x,t) = 2\,\mathrm{e}^{-(\frac{\pi}{2})^2 t} \operatorname{sen}(2\pi x)$.

9. Desarróllese un método de diferencias finitas para resolver la EDP parabólica no lineal $u_{xx} = u\, u_t$.

10. La *ecuación del calor en dos dimensiones espaciales* es $u_t = \beta(u_{xx} + u_{yy})$, que modela la distribución de temperatura sobre la superficie de una placa calentada. Sin embargo, más que caracterizar la distribución en un estado estacionario, como se hace para la ecuación de Laplace, esta ecuación ofrece un medio para calentar la distribución de temperatura de la placa conforme cambia con el tiempo. Obténgase una solución explícita sustituyendo en la EDP las aproximaciones por diferencias finitas, tal y como se hace para la ecuación del calor en una dimensión espacial. Demuéstrese que son necesarios cinco puntos en un instante anterior para calcular cada nuevo valor de u_{ij}.

11. *Condiciones aisladas para la ecuación del calor.* La forma de las condiciones de contorno depende del hecho físico que se describa. Por ejemplo, si se mantiene aislado un extremo de la varilla en la ecuación del calor, entonces la derivada parcial u_x es cero en ese extremo; es decir, $u_x(0,t) = 0$ o $u_x(\ell,t) = 0$. Desarróllese el método explícito para la ecuación del calor cuando la temperatura está dada en un extremo ($x = 0$), mientras que el otro extremo ($x = \ell$) está aislado. (Se recomienda, para la condición de contorno que se dé en la derivada, añadir un punto ficticio a la malla; habitualmente se extiende la malla para incluir un punto en cada paso de tiempo.)

12. Utilícese el método de diferencias finitas para calcular las tres primeras filas de la solución aproximada de la siguiente ecuación de ondas

$$\begin{cases} u_{tt} = 4\,u_{xx}, \quad 0 < x < 1, \quad t > 0, \\ u(0,t) = u(1,t) = 0, \quad t > 0, \\ u(x,0) = \begin{cases} \frac{5x}{2}, & \text{para } 0 \le x \le \frac{3}{5}, \\ \frac{15}{4}(1-x), & \text{para } \frac{3}{5} \le x \le 1, \end{cases} \\ u_t(x,0) = 0, \quad 0 \le x \le 1. \end{cases}$$

Tómese $h = 0.2$ y $k = 0.1$. Realícense las operaciones a mano (o con calculadora).

13. En la ecuación $u_{tt} = 9\,u_{xx}$, ¿qué relación debe existir entre h y k para que la ecuación en diferencias que se obtenga esté dada por $u_{i,j+1} = u_{i-1,j} + u_{i+1,j} - u_{i,j-1}$?

14. ¿Qué dificultad puede aparecer cuando se intenta utilizar el método de diferencias finitas para resolver $u_{tt} = 4\,u_{xx}$ tomando $h = 0.03$ y $k = 0.02$? ¿Cómo se podría solventar dicha dificultad? Propóngase y llévese a cabo una solución para solventar dicha dificultad.

15. *Método implícito para la ecuación de ondas.* Para la ecuación de ondas, como en el caso de la ecuación del calor, los métodos implícitos tienen ventajas de estabilidad. Un esquema implícito simple resulta de reemplazar u_{tt} por una fórmula de diferencias centradas en el i-ésimo paso de espacio y u_{xx} por el valor medio de una diferencia centrada en los $(j-1)$-ésimo y $(j+1)$-ésimo pasos de tiempo. Desarróllese dicho esquema.

16. Sea la ecuación de Poisson

$$\begin{cases} u_{xx} + u_{yy} = f(x,y), \quad 0 \le x \le 1, \quad 0 \le y \le 1, \\ u(x,0) = x, \quad u(x,1) = 1, \quad 0 \le x \le 1, \\ u(0,y) = y, \quad u(1,y) = 1, \quad 0 \le y \le 1, \end{cases} \qquad f(x,y) = \begin{cases} 1, & \text{para } (x,y) = \left(\frac{1}{2}, \frac{2}{3}\right), \\ 0, & \text{en otro caso.} \end{cases}$$

Constrúyase una malla apropiada y resuélvase la ecuación mediante el MEF.



Complementos

Complemento A

Valores propios y vectores propios

A.1. Introducción

La ecuación $A\mathbf{v} = \lambda\mathbf{v}$, donde A es una matriz cuadrada de orden n, λ es un valor propio de A y $\mathbf{v}$ el vector propio asociado a λ, se puede interpretar de forma mucho más general. La multiplicación $A\mathbf{v}$ es una operación matemática que se puede ver como la matriz A actuando sobre el operando $\mathbf{v}$, de manera que la ecuación anterior se puede generalizar a cualquier operación matemática de la forma $L\mathbf{v} = \lambda\mathbf{v}$, donde L es un operador que puede representar la multiplicación por una matriz, diferenciación, integración, etc., $\mathbf{v}$ es un vector o una función y λ es un escalar. Por ejemplo, si L representa la derivada segunda con respecto a x, y es una función de x y λ es una constante, la ecuación $L\mathbf{v} = \lambda\mathbf{v}$ puede tomar la forma $y''(x) = \lambda y(x)$. Luego existe cierta conexión entre los problemas de valores propios que involucran EDO y problemas de valores propios que involucran matrices.

La ecuación $L\mathbf{v} = \lambda\mathbf{v}$ es una forma general de un problema de valores propios, donde λ es el valor propio asociado al operador L y $\mathbf{v}$ es el vector propio o función propia asociada al valor propio λ y al operador L.

Los problemas de valores propios son un tipo especial de problemas de contorno con EDO que tienen especial importancia en ingeniería. Por ejemplo, aparecen en problemas que implican vibraciones, en problemas de la mecánica de fluidos, en problemas de elasticidad, en problemas de dinámica espacial, así como en otros muchos problemas. También se utilizan en otros diversos contextos de la ingeniería diferentes del de los problemas de contorno con EDO. Además, es importante destacar que los valores propios de una matriz proporcionan información útil sobre algunas propiedades de los cálculos numéricos que involucran a la matriz. Recordemos que la convergencia de los métodos iterativos para resolver sistemas lineales dependía de los valores propios de sus matrices de iteración, y que la velocidad con que estos métodos convergen depende de la magnitud de sus valores propios.

Los métodos numéricos que aproximan los valores propios de una matriz no suelen calcular el polinomio característico, sino que actúan directamente sobre la matriz. En muchos problemas solo se necesita conocer el valor propio dominante (el de mayor módulo) y tal vez su vector propio asociado. Para tales problemas el método más conocido que se utiliza es el *método de la potencia*. Con ligeras modificaciones de este método también se pueden calcular el valor propio de menor módulo y los valores propios intermedios.

A.2. El método de la potencia

El método de la potencia es un método iterativo que aproxima el mayor valor propio en módulo. El vector propio asociado al valor propio se obtiene como parte del método. Frecuentemente, también se emplea el método de la potencia para calcular un vector propio asociado a un valor propio ya calculado por otros medios.

Si A es una matriz cuadrada de orden n, para poder aplicar el método de la potencia, tiene que tener n valores propios $\lambda_1, \lambda_2, \ldots, \lambda_n$ con un conjunto linealmente independiente de vectores propios asociados $\{\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_n}\}$ y tales que

$$|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \cdots \geq |\lambda_n| \,.$$

Como los vectores $\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_n}$ son linealmente independientes, podemos escribir un vector cualquiera

$\mathbf{v} \neq 0$ de $\mathbb{R}^n$ como

$$\mathbf{v} = c_1\mathbf{v_1} + c_2\mathbf{v_2} + \cdots + c_n\mathbf{v_n}, \tag{A.1}$$

donde $c_i \neq 0$ son escalares para todo $i = 1, 2, \ldots, n$. Sea $\mathbf{w_0} = \mathbf{v}$. Multiplicando (A.1) por A, obtenemos

$$A\mathbf{w_0} = c_1 A\mathbf{v_1} + c_2 A\mathbf{v_2} + \cdots + c_n A\mathbf{v_n} = \lambda_1 c_1 \mathbf{w_1},$$

donde $\mathbf{w_1} = \mathbf{v_1} + \frac{c_2}{c_1}\frac{\lambda_2}{\lambda_1}\mathbf{v_2} + \cdots + \frac{c_n}{c_1}\frac{\lambda_n}{\lambda_1}\mathbf{v_n}$, puesto que $A\mathbf{v_j} = \lambda_j \mathbf{v_j}$, para todo $j = 1, 2, \ldots, n$, por ser $\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_n}$ vectores propios. Si multiplicamos ahora $\mathbf{w_1}$ por A, tenemos

$$A\mathbf{w_1} = \lambda_1\mathbf{v_1} + \frac{c_2}{c_1}\frac{\lambda_2^2}{\lambda_1}\mathbf{v_2} + \cdots + \frac{c_n}{c_1}\frac{\lambda_n^2}{\lambda_1}\mathbf{v_n} = \lambda_1\left(\mathbf{v_1} + \frac{c_2}{c_1}\frac{\lambda_2^2}{\lambda_1^2}\mathbf{v_2} + \cdots + \frac{c_n}{c_1}\frac{\lambda_n^2}{\lambda_1^2}\mathbf{v_n}\right) = \lambda_1\mathbf{w_2} \tag{A.2}$$

con $\mathbf{w_2} = \mathbf{v_1} + \frac{c_2}{c_1}\frac{\lambda_2^2}{\lambda_1^2}\mathbf{v_2} + \cdots + \frac{c_n}{c_1}\frac{\lambda_n^2}{\lambda_1^2}\mathbf{v_n}$. Multiplicando a continuación (A.2) por A, da

$$A\mathbf{w_2} = \lambda_1\mathbf{v_1} + \frac{c_2}{c_1}\frac{\lambda_2^3}{\lambda_1^2}\mathbf{v_2} + \cdots + \frac{c_n}{c_1}\frac{\lambda_n^3}{\lambda_1^2}\mathbf{v_n} = \lambda_1\mathbf{w_3},$$

donde $\mathbf{w_3} = \mathbf{v_1} + \frac{c_2}{c_1}\frac{\lambda_2^3}{\lambda_1^3}\mathbf{v_2} + \cdots + \frac{c_n}{c_1}\frac{\lambda_n^3}{\lambda_1^3}\mathbf{v_n}$. Es fácil ver que iterando sucesivamente llegamos a

$$\mathbf{w_k} = \mathbf{v_1} + \frac{c_2}{c_1}\frac{\lambda_2^k}{\lambda_1^k}\mathbf{v_2} + \cdots + \frac{c_n}{c_1}\frac{\lambda_n^k}{\lambda_1^k}\mathbf{v_n}. \tag{A.3}$$

Como λ_1 es el mayor valor propio en módulo, entonces $\left|\frac{\lambda_j}{\lambda_1}\right| < 1$, para todo $j = 2, 3, \ldots, n$, de manera que el segundo miembro de (A.3) tiende a $\mathbf{v_1}$ cuando $k \to \infty$. Por lo tanto, si $k \to \infty$, obtenemos

$$\mathbf{w_k} \to \mathbf{v_1} \qquad \text{y} \qquad A\mathbf{w_k} \to \lambda_1\mathbf{v_1}.$$

Cuando se implementa el método de la potencia, el vector $\mathbf{w_k}$ se normaliza en cada paso dividiendo las componentes del vector por el móldulo de la mayor componente (véase (A.1) a través de (A.3)). Esto hace que la mayor componente del vector sea 1. Como consecuencia de este escalado, el método de la potencia conduce al valor propio y al vector propio asociado.

El método de la potencia se aplica generalmente como sigue. Se empieza con una estimación inicial $\mathbf{w_0}$ de $\mathbf{v}$. Normalmente se elige esta estimación inicial como $\mathbf{w_0} = (1, 1, \ldots, 1)^T$, de manera que la norma infinito $\|\mathbf{w_0}\|_\infty$ es 1. Entonces se genera recursivamente la sucesión $\{\mathbf{w_k}\}$, a partir de

$$\begin{aligned} \mathbf{z_{k-1}} &= A\mathbf{w_{k-1}}, \\ \mathbf{w_k} &= \frac{1}{d_k}\mathbf{z_{k-1}}, \end{aligned}$$

donde d_k es la mayor componente en módulo de $\mathbf{z_{k-1}}$. Si el método converge, el valor final de d_k es el valor propio buscado y el valor final de $\mathbf{w_k}$ es el vector propio asociado. Esto es,

$$\lim_{k\to\infty} \mathbf{w_k} = \mathbf{v_1} \qquad \text{y} \qquad \lim_{k\to\infty} d_k = \lambda_1.$$

El proceso iterativo se termina cuando la norma infinito $\|\mathbf{w_k} - \mathbf{w_{k-1}}\|_\infty$ es menor que una tolerancia especificada.

Ejemplo. La matriz

$$A = \begin{pmatrix} -9 & 14 & 4 \\ -7 & 12 & 4 \\ 0 & 0 & 1 \end{pmatrix}$$

tiene valores propios $\lambda_1 = 5$, $\lambda_2 = 1$ y $\lambda_3 = -2$, así que el método de la potencia converge. Comenzamos con $\mathbf{w_0} = (1, 1, 1)^T$. Aplicando el método de la potencia se calcula el vector $\mathbf{w_1}$ a partir de $A\mathbf{w_0}$ y normalizando:

$$A\mathbf{w_0} = \begin{pmatrix} -9 & 14 & 4 \\ -7 & 12 & 4 \\ 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 9 \\ 9 \\ 1 \end{pmatrix} = 9\begin{pmatrix} 1 \\ 1 \\ \frac{1}{9} \end{pmatrix} \quad \Rightarrow \quad d_1 = 9, \quad \mathbf{w_1} = \begin{pmatrix} 1 \\ 1 \\ \frac{1}{9} \end{pmatrix}.$$

La siguiente iteración será:

$$A\mathbf{w_1} = \begin{pmatrix} -9 & 14 & 4 \\ -7 & 12 & 4 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ \frac{1}{9} \end{pmatrix} = \frac{49}{9} \begin{pmatrix} 1 \\ 1 \\ \frac{1}{49} \end{pmatrix} \quad \Rightarrow \quad d_2 = \frac{49}{9}, \quad \mathbf{w_2} = \begin{pmatrix} 1 \\ 1 \\ \frac{1}{49} \end{pmatrix}.$$

Después de diez iteraciones, la sucesión de vectores converge al vector $\mathbf{v} = (1, 1, 1.02 \times 10^{-8})^T$ y la sucesión $\{d_k\}$ de constantes converge a $\lambda = 5.000000205$. En [16] se encuentra una tabla que resume los cálculos. □

Notas sobre la convergencia del método de la potencia

- Generalmente el método de la potencia converge muy lentamente, a menos que el vector inicial esté próximo al vector $\mathbf{v_1}$. Puede surgir un problema cuando el vector inicial es tal que el valor de c_1 en (A.1) es cero. Esto significa que $\mathbf{v}$ no tiene componentes en la dirección del vector $\mathbf{v_1}$. Teóricamente el método fallará. En la práctica, sin embargo, el método puede converger (muy lentamente) a causa de la acumulación de errores de redondeo durante la repetida multiplicación con la matriz A, que proporcionará componentes en la dirección de $\mathbf{v_1}$. (Por una vez podemos decir que los errores de redondeo ayudan.)
- En realidad no es necesario que la matriz A tenga n valores propios distintos para que el método de la potencia converja. Si la matriz tiene un único valor propio dominante λ_1 de multiplicidad $m > 1$ y $\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_m}$ son vectores propios linealmente independientes asociados a λ_1, el método seguirá convergiendo a λ_1.
- El método de la potencia tiene la desventaja de que no se sabe de antemano si la matriz tiene un único valor propio dominante, en cuyo caso el método podría no converger.

A.3. El método de la potencia con desplazamiento

La mayor desventaja del método de la potencia es que su velocidad de convergencia es lenta cuando la razón de predominio $r = \left|\frac{\lambda_2}{\lambda_1}\right|$ de los dos mayores valores propios en módulo está próxima a 1. Sin embargo, la velocidad de convergencia se puede acelerar mediante varias estrategias. La estrategia más simple consiste en utilizar el método de la potencia con desplazamiento. Sabemos que A y $A - qI$ tienen el mismo conjunto de vectores propios, y para cada valor propio λ de A, tenemos el valor propio $\lambda - q$ de $A - qI$. Esto es,

$$(A - qI)\mathbf{v} = A\mathbf{v} - q\mathbf{v} = \lambda\mathbf{v} - q\mathbf{v} = (\lambda - q)\mathbf{v}.$$

Por lo tanto, si restamos q de todos los elementos diagonales de A, cada valor propio se reduce por el mismo factor y los vectores propios no cambian. El método de la potencia se puede utilizar ahora así:

$$\begin{aligned} \mathbf{z_k} &= (A - qI)\mathbf{w_k}, \\ \mathbf{w_{k+1}} &= \frac{1}{d_{k+1}}\mathbf{z_k}. \end{aligned}$$

Por ejemplo, si 20 y -18 son los dos valores propios mayores en módulo, la razón de predominio $r = 0.9$ está próxima a 1 y la velocidad de convergencia será entonces lenta. Sin embargo si elegimos $q = \frac{1}{2}(10 + 18) = 14$ entonces los nuevos valores propios son 6 y -32, de manera que la razón de predominio es $r = 0.1875$ y se obtendrá mayor velocidad de convergencia.

Por supuesto que la elección de q es difícil, a menos que sepamos a priori una estimación de los valores propios. Una forma de obtener una estimación de los valores propios es utilizando el siguiente resultado, conocido como **teorema de los círculos de Gerschgorin**:

> Sean A una matriz cuadrada de orden n y C_i el círculo del plano complejo con centro a_{ii} y radio $r_i = \sum_{j=1,\, j\neq i}^{n} |a_{ij}|$, es decir, $C_i = \{z \in \mathbb{C};\ |z - a_{ii}| \leq r_i\}$. Entonces, los valores propios de A están contenidos en $D = \bigcup_{i=1}^{n} C_i$. Además, si la unión de k de estos círculos no se corta con los restantes $n - k$, entonces dicha unión contiene precisamente k valores propios (contando las multiplicidades).

Ejemplo. Dada la matriz

$$A = \begin{pmatrix} 2 & 3 & 0 \\ 3 & 8 & -2 \\ 0 & -2 & -4 \end{pmatrix}$$

observamos que todos sus valores propios son reales, puesto que A es simétrica y los círculos de Gerschgorin son:

C_1 es el círculo con centro en $(2, 0)$ y radio $= 3 + 0 = 3$,

C_2 es el círculo con centro en $(8, 0)$ y radio $= 3 + |-2| = 5$,

C_3 es el círculo con centro en $(-4, 0)$ y radio $= 0 + |-2| = 2$.

La unión de estos círculos es

$$D = [-1, 5] \cup [3, 13] \cup [-6, -2] = [-6, -2] \cup [-1, 13].$$

Como C_1 y C_2 son disjuntos con C_3, véase la figura A.1, hay dos valores propios en $C_1 \cup C_2$ y uno en C_3. En efecto, los valores propios de A son: $\lambda_1 = 9.4969$, $\lambda_2 = 0.8684$ y $\lambda_3 = -4.3653$. □

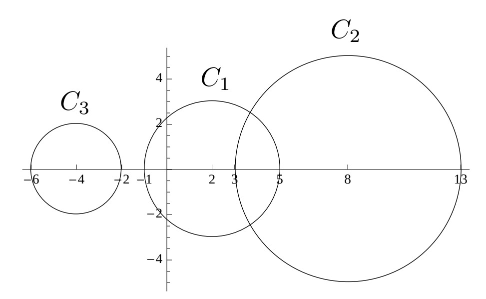


Figura A.1: Círculos de Gerschgorin para la matriz A del ejemplo.

Notemos que no hay garantía de que un círculo contenga valores propios a menos que éste esté aislado de los otros. El teorema de los círculos de Gerschgorin se puede aplicar para obtener una conjetura preliminar del *desplazamiento*, como mostramos en el siguiente ejemplo.

Ejemplo. En el ejemplo anterior hemos visto que los valores propios de la matriz A están en $D = [-1, 5] \cup [3, 13] \cup [-6, -2] = [-6, -2] \cup [-1, 13]$, de manera que para calcular el mayor valor propio en módulo, podemos utilizar el método de la potencia con desplazamiento poniendo el valor del desplazamiento $q = 13$. □

Comentario adicional. ▷ El método de la potencia con desplazamiento permite calcular el valor propio más cercano a un número dado, y para que su aplicación sea efectiva necesitamos cierto conocimiento a priori de la situación de los valores propios de la matriz, que puede conseguirse inspeccionando los círculos de Gerschgorin.

A.4. El método de la potencia inversa

También podemos obtener el menor valor propio en módulo de forma similar al de mayor módulo, mediante el método de la potencia inversa. Si la matriz A es ivertible y λ es un valor propio de A y $\mathbf{v}$ su vector propio asociado, entonces $\frac{1}{\lambda}$ es un valor propio de A^{-1} asociado al mismo vector propio $\mathbf{v}$, ya que $A^{-1}\mathbf{v} = \frac{1}{\lambda}\mathbf{v}$ si $A\mathbf{v} = \lambda\mathbf{v}$. Aplicando el método de la potencia a A^{-1} podemos obtener una aproximación al menor valor propio de A en módulo (siempre que exista).

Eligiendo un vector inicial adecuado $\mathbf{v} \neq 0$, en el primer paso obtenemos $\mathbf{u}$ tal que $A\mathbf{u} = \mathbf{v}$, que es equivalente a $\mathbf{u} = A^{-1}\mathbf{v}$. Obviamente la matriz inversa A^{-1} se tiene que calcular con anterioridad. En la práctica, sin embargo, como calcular la inversa de una matriz es computacionalmente ineficiente y no deseable, se suele resolver el sistema lineal $A\mathbf{u} = \mathbf{v}$ (utilizando habitualmente un método de factorización LU). A partir de aquí funciona todo igual que en el método de la potencia.

A.5. Cálculo de todos los valores propios

Una vez que el mayor (o menor) valor propio en módulo es conocido, se puede utilizar el método de la potencia con desplazamiento para encontrar los otros valores propios. Este método utiliza la siguiente propiedad importante de las matrices y sus valores propios:

> Dado $A\mathbf{v} = \lambda\mathbf{v}$, si λ_1 es el mayor (o menor) valor propio en módulo de A, obtenido mediante el método de la potencia (o el método de la potencia inversa), entonces los nuevos valores propios de la matriz *desplazada* $A - \lambda_1 I$ son $0, \lambda_2 - \lambda_1, \lambda_3 - \lambda_1, \ldots, \lambda_n - \lambda_1$.

Esto se puede ver fácilmente porque los valores propios de $A - \lambda_1 I$ cumplen

$$(A - \lambda_1 I)\mathbf{v} = \mu\mathbf{v} \tag{A.4}$$

donde μ es el valor propio de la matriz desplazada $A - \lambda_1 I$. Pero, como $A\mathbf{v} = \lambda\mathbf{v}$, entonces (A.4) es $(\lambda - \lambda_1)\mathbf{v} = \mu\mathbf{v}$, donde $\lambda = \lambda_1, \lambda_2, \ldots, \lambda_n$. Por lo tanto, los valores propios de la matriz desplazada $A - \lambda_1 I$ son $\mu = 0, \lambda_2 - \lambda_1, \lambda_3 - \lambda_1, \ldots, \lambda_n - \lambda_1$, y los vectores propios son los mismos que los de A. Ahora si se aplica el método de la potencia a la matriz desplazada $A - \lambda_1 I$, después de haberse aplicado a A para calcular λ_1, se puede calcular el mayor valor propio en módulo μ_k de la matriz desplazada. Entonces, el valor propio λ_k se puede determinar como $\lambda_k = \mu_k + \lambda_1$. El resto de valores propios se pueden calcular repitiendo este proceso $k - 2$ veces, donde cada vez la matriz desplazada es $A - \lambda_k I$ y λ_k el valor propio obtenido a partir del desplazamiento anterior.

COMENTARIO ADICIONAL. ▷ La combinación del teorema de Gerschgorin, el método de la potencia inversa y el método de la potencia con desplazamiento puede ser útil para calcular los valores propios de una matriz. Por ejemplo, si hay un círculo de Gerschgorin con un valor propio asociado, podemos llevar a cabo un desplazamiento q similar al centro de dicho círculo para que el valor propio correspondiente de la matriz desplazada esté próximo a cero y ahora aplicar el método de la potencia inversa para calcularlo. Después basta con invertir el valor obtenido y deshacer el desplazamiento.

Notas finales

- Una vez calculado el valor propio dominante, también se puede utilizar el método de la potencia para determinar los demás valores propios de una matriz mediante *técnicas de deflación*, que consisten en construir una nueva matriz B cuyos valores propios son los mismos que los de A, excepto el dominante, que se sustituye por 0 como valor propio de B. Por tanto, si λ_2 «domina» a $\lambda_3, \lambda_4, \ldots, \lambda_n$, podemos aplicar de nuevo el método de la potencia a la matriz B para calcular λ_2, y así sucesivamente. Véase [7] para el método concreto de Wielandt.

- Hay una gran variedad de métodos para calcular los valores propios de una matriz, la mayoría se base en un proceso de dos etapas: en la primera se transforma la matriz original en una matriz más simple que conserve todos los valores propios de la matriz original, y en la segunda se determinan los valores propios mediante un método iterativo. Muchos de estos métodos están diseñados para tipos especiales de matrices. Cuando se buscan los valores propios de una matriz general cualquiera, los métodos LR de Rutishauser y QR de Francis son los más importantes. Aunque el método QR es menos eficiente, frecuentemente es el preferido porque es más estable. En [15] se pueden ver varios métodos.

A.6. Ejercicios

1. Calcúlese el mayor y el menor valor propio en módulo de las siguiente matrices:

$$A = \begin{pmatrix} 2 & 2 & -1 \\ -5 & 9 & -3 \\ -4 & 4 & 1 \end{pmatrix}, \qquad B = \begin{pmatrix} 1 & 0 & 0 \\ 2 & -1 & 2 \\ 4 & -4 & 5 \end{pmatrix}, \qquad C = \begin{pmatrix} 5 & -2 & 1 \\ 3 & 0 & 1 \\ 0 & 0 & 2 \end{pmatrix}.$$

2. Utilícese el teorema de los círculos de Gerschgorin para hallar un desplazamiento adecuado para calcular el mayor valor propio en módulo de la matriz

$$A = \begin{pmatrix} 5 & 0 & 1 & -1 \\ 0 & 2 & 0 & -\frac{1}{2} \\ 0 & 1 & -1 & 1 \\ -1 & -1 & 0 & 0 \end{pmatrix}.$$

   Compárense los números de iteraciones que utilizan el método de la potencia y el método de la potencia con desplazamiento.

3. Verifíquese que el método de la potencia no converge al mayor valor propio en módulo de la matriz

$$A = \begin{pmatrix} \frac{1}{3} & \frac{2}{3} & 2 & 3 \\ 1 & 0 & -1 & 2 \\ 0 & 0 & -\frac{5}{3} & -\frac{2}{3} \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

   Explíquese por qué.

4. Sea la matriz

$$A = \begin{pmatrix} 0 & 0 & 2 & 4 \\ \frac{1}{2} & 0 & 0 & 0 \\ 0 & \frac{1}{4} & 0 & 0 \\ 0 & 0 & \frac{1}{8} & 0 \end{pmatrix}.$$

   a) Determínese una región del plano complejo que contenga a todos los valores propios de A.
   b) Encuéntrense el mayor valor propio en módulo de A y su vector propio asociado.
   c) Hállense el resto de valores propios de A.
   d) Calcúlense los valores propios de A a partir de su polinomio característico y aplicando el método de Newton.

5. Dada la matriz

$$A = \begin{pmatrix} 2 & 0 & -1 \\ -2 & -10 & 0 \\ -1 & -1 & 4 \end{pmatrix},$$

   utilícese el método de la potencia para calcular el radio espectral de A con la norma infinito.

Complemento B

Introducción a las ecuaciones diferenciales en derivadas parciales

B.1. Introducción

Actualmente la teoría de las ecuaciones diferenciales en derivadas parciales (abreviadamente, EDP) es uno de los campos más importantes de estudio de las matemáticas, ya que estas ecuaciones aparecen frecuentemente en muchas ramas de la física y de la ingeniería, así como de otras ciencias.

En muchas formulaciones de modelos matemáticos se utilizan derivadas parciales para representar cantidades físicas. Estas derivadas siempre dependen de más de una variable independiente, generalmente las variables espacio $x, y, \ldots$ y la variable tiempo t. Tales formulaciones tienen una o más variables dependientes, que son funciones desconocidas de las variables independientes. Las ecuaciones resultantes se llaman ecuaciones diferenciales en derivadas parciales, que junto con condiciones iniciales y de contorno, representan fenómenos físicos.

En este capítulo daremos una brevísima introducción a las EDP. Éste es un tema tan amplio que necesitaría por sí solo de un texto completo para una simple introducción. Mostraremos los aspectos más básicos de las EDP, limitándonos al estudio de las EDP de segundo orden, ya que cubren las más características e importantes ecuaciones de la física matemática: la ecuación de ondas, la ecuación del calor y la ecuación de Laplace. Introduciremos meramente ciertos conceptos fundamentales y algún método básico de resolución.

B.2. EDP y sus soluciones

Recordemos en primer lugar que una EDP es una ecuación diferencial que contiene derivadas parciales de una o más variables dependientes respecto a una o más variables independientes. Por ejemplo, si u es una función de las variables independientes x e y, toda ecuación que contenga $\frac{\partial u}{\partial x}$ o $\frac{\partial u}{\partial y}$, o derivadas de orden superior, y también u, x o y, es una EDP. Algunos ejemplos importantes de las ciencias físicas son:

i) $\dfrac{\partial^2 u}{\partial t^2} = \alpha^2 \dfrac{\partial^2 u}{\partial x^2}$ es la *ecuación unidimensional de ondas*,

ii) $\dfrac{\partial u}{\partial t} = \beta \dfrac{\partial^2 u}{\partial x^2}$ es la *ecuación unidimensional del calor*,

iii) $\dfrac{\partial^2 u}{\partial x^2} + \dfrac{\partial^2 u}{\partial y^2} = 0$ es la *ecuación bidimensional de Laplace*,

iv) $\dfrac{\partial^2 u}{\partial x^2} + \dfrac{\partial^2 u}{\partial y^2} = f(x, y)$ es la *ecuación bidimensional de Poisson*,

v) $\dfrac{\partial^2 u}{\partial x^2} + \dfrac{\partial^2 u}{\partial y^2} + \dfrac{\partial^2 u}{\partial z^2} = 0$ es la *ecuación tridimensional de Laplace*.

En las dos primeras EDP, la función incógnita u es una función de la coordenada x y del tiempo t: $u = u(x,t)$. En las tercera y cuarta EDP, la función u depende de las dos coordenadas de un punto en el plano, y en última la función depende de las tres coordenadas de un punto en el espacio.

Llamamos **orden** de una EDP al que tiene la derivada de mayor orden que aparece en la ecuación. En los ejemplos anteriores todas las EDP son de segundo orden.

Al igual que para las EDO, hay varios tipos estándares de EDP importantes que se dan con frecuencia en los modelos matemáticos de los sistemas físicos, y cuyas soluciones se pueden expresar mediante funciones elementales. Una **solución** de una EDP es una relación explícita o implícita entre las variables, relación que no contiene derivadas y que satisface identicamente la ecuación. En determinados casos muy simples se puede obtener una solución inmediatamente.

EJEMPLO. Si consideramos la EDP de primer orden

$$\frac{\partial u}{\partial x} = x^2 + y^2,$$

en la que u es la variable dependiente y x e y son las variables independientes, la solución es

$$u(x,y) = \int (x^2 + y^2)\,\partial x + \phi(y),$$

donde $\int (x^2 + y^2)\,\partial x$ indica una «integración parcial» respecto a x, manteniendo y constante, y ϕ es una función arbitraria que depende solo de y. Por tanto, la solución de la EDP es

$$u(x,y) = \frac{x^3}{3} + xy^2 + \phi(y). \quad \square$$

EJEMPLO. Sea la EDP de segundo orden

$$\frac{\partial^2 u}{\partial y\,\partial x} = x^3 - y.$$

Primero la escribimos en la forma

$$\frac{\partial}{\partial y}\left(\frac{\partial u}{\partial x}\right) = x^3 - y$$

y la integramos parcialmente respecto a y, manteniendo x constante, con lo que obtenemos

$$\frac{\partial u}{\partial x} = x^3 y - \frac{1}{2}y^2 + \phi(x),$$

donde ϕ es una función arbitraria de x. Integramos ahora este resultado parcialmente respecto a x, manteniendo y constante, con lo que obtenemos la solución de la EDP

$$u(x,y) = \frac{1}{4}x^4 y - \frac{1}{2}xy^2 + f(x) + g(y),$$

donde f es una función arbitraria de x que está definida por $f(x) = \int \phi(x)\,dx$ y g es una función de y también arbitraria. $\square$

Como resultado de estos dos sencillos ejemplos observamos que mientras que las EDO tienen soluciones que dependen de *constantes* arbitrarias, las soluciones de las EDP dependen de *funciones* arbitrarias. En particular, observamos que la solución de la EDP de *primer orden* contiene *una* función arbitraria y que la solución de la EDP de *segundo orden* contiene *dos* funciones arbitrarias. En general, la solución de una EDP contiene un cierto número de funciones arbitrarias, a menudo n funciones para una ecuación de orden n.

NOTA. La naturaleza de un problema matemático típico que se refiere a una EDP y que se origina en la formulación matemática de algún problema físico consta no solo de la propia ecuación diferencial, sino que contiene también ciertas condiciones suplementarias (denominadas *condiciones de contorno*, *condiciones iniciales* o ambas). El número y la naturaleza de estas condiciones depende de la naturaleza del problema físico que ha originado el problema matemático. La solución del problema ha de satisfacer tanto la ecuación diferencial como las condiciones suplementarias. Con otras palabras, la solución del problema completo (ecuación diferencial más condiciones) es una *solución particular* de la EDP del problema.

B.3. EDP lineales de segundo orden

Consideramos brevemente la clase de EDP que aparecen con más frecuencia, que es la de las denominadas EDP lineales de segundo orden. Una EDP lineal de segundo orden con dos variables independientes x e y es una ecuación de la forma:

$$Au_{xx} + Bu_{xy} + Cu_{yy} + Du_x + Eu_y + Fu = G, \tag{B.1}$$

donde A, B, C, D, E, F y G son funciones de x e y. Supongamos que A, B y C no se anulan simultaneamente. Si $G \equiv 0$, la EDP (B.1) se reduce a

$$Au_{xx} + Bu_{xy} + Cu_{yy} + Du_x + Eu_y + Fu = 0. \tag{B.2}$$

Para esta clase de EDP podemos enunciar el siguiente **teorema fundamental** con respecto a sus soluciones:

> Sean n soluciones $u_1, u_2, \dots, u_n$ de la EDP (B.2) en una región $\mathcal{R}$ del plano XY. La combinación lineal $c_1u_1 + c_2u_2 + \cdots + c_nu_n$, donde $c_1, c_2, \dots, c_n$ son constantes arbitrarias, es también una solución de la EDP (B.2) en $\mathcal{R}$.

Consideramos ahora una clase especial importante de EDP lineales de segundo orden, las denominadas EDP lineales *homogéneas* de segundo orden *con coeficientes constantes*. Una ecuación de esta clase es de la forma

$$au_{xx} + bu_{xy} + cu_{yy} = 0 \quad \text{con } a, b \text{ y } c \text{ contantes.} \tag{B.3}$$

La palabra *homogénea* hace referencia aquí al hecho de que todos los términos en (B.3) contienen derivadas del *mismo* orden (el segundo).

Buscamos ahora una solución de (B.3) en la forma:

$$u(x, y) = f(y + \lambda x), \tag{B.4}$$

donde f es una función arbitraria de su argumento y λ es una constante. Derivando (B.4), obtenemos

$$u_{xx}(x, y) = \lambda^2 f''(y + \lambda x), \qquad u_{xy}(x, y) = \lambda f''(y + \lambda x), \qquad u_{yy}(x, y) = f''(y + \lambda x),$$

y sustituyéndolas en (B.3) da

$$a\lambda^2 f''(y + \lambda x) + b\lambda f''(y + \lambda x) + cf''(y + \lambda x) = 0 \qquad \Rightarrow \qquad f''(y + \lambda x)[a\lambda^2 + b\lambda + c] = 0,$$

de manera que $u(x, y) = f(y + \lambda x)$ es una solución de (B.3) si λ satisface la ecuación de segundo grado

$$a\lambda^2 + b\lambda + c = 0,$$

que llamaremos **ecuación condicionante**.

Consideramos ahora los siguientes cuatro casos de la EDP (B.3), véase [21]:

1. $a \neq 0$ y las raíces de la ecuación condicionante diferentes,
2. $a \neq 0$ y las raíces de la ecuación condicionante iguales,
3. $a = 0$, $b \neq 0$,
4. $a = 0$, $b = 0$ y $c \neq 0$.

En el caso 1, la EDP (B.3) posee dos soluciones $u(x, y) = f(y + \lambda_1 x)$ y $u(x, y) = g(y + \lambda_2 x)$, donde f y g son dos funciones arbitrarias de sus respectivos argumentos y λ_1 y λ_2 son las raíces distintas de la ecuación condicionante. Aplicando el teorema fundamental anterior se deduce que

$$u(x, y) = f(y + \lambda_1 x) + g(y + \lambda_2 x)$$

es una solución de la EDP (B.3).

En el caso 2, la EDP (B.3) posee la solución $u(x, y) = f(y + \lambda_1 x)$, donde f es una función arbitraria de su argumento y λ_1 es la raíz doble de la ecuación condicionante. Además, se puede comprobar fácilmente que en este caso la EDP (B.3) posee también la solución $u(x, y) = xg(y + \lambda_1 x)$, donde g es una función arbitraria de su argumento. Según el teorema fundamental anterior vemos que

$$u(x, y) = f(y + \lambda_1 x) + xg(y + \lambda_1 x)$$

es una solución de la EDP (B.3).

En el caso 3, la ecuación condicionante se reduce a $b\lambda + c = 0$, por lo que tiene una única raíz $\lambda_1 = -c/b$. La EDP (B.3) posee la solución $u(x, y) = f(y + \lambda_1 x)$, donde f es una función arbitraria de su argumento. Se puede comprobar además que en este caso $g(x)$, donde g es una función arbitraria de x solamente, es también una solución de la EDP (B.3). Según el teorema fundamental anterior tenemos entonces que

$$u(x, y) = f(y + \lambda_1 x) + g(x)$$

es una solución de la EDP (B.3).

Finalmente, en el caso 4, la ecuación condicionante se reduce a $c = 0$, lo que es imposible. Vemos que en este caso no existen soluciones de la forma (B.4). No obstante, la EDP es ahora simplemente

$$c\,\frac{\partial^2 u}{\partial y^2} = 0 \qquad \text{o} \qquad \frac{\partial}{\partial y}\left(\frac{\partial u}{\partial y}\right) = 0.$$

Integrando parcialmente respecto a y dos veces, obtenemos $u(x, y) = f(x) + yg(x)$, donde f y g son funciones arbitrarias de x solamente. Por tanto, en este caso,

$$u(x, y) = f(x) + yg(x)$$

es una solución de la EDP (B.3).

Notemos que toda EDP con coeficientes constantes de la forma (B.3) está dentro de una y solo una de las cuatro categorías cubiertas por los casos 1 al 4.

EJEMPLO. Sea

$$u_{xx} - 5u_{xy} + 6u_{yy} = 0.$$

La ecuación condicionante correspondiente a esta EDP es $\lambda^2 - 5\lambda + 6 = 0$, que posee dos raíces diferentes $\lambda_1 = 2$ y $\lambda_3 = 3$. Estamos entonces en el caso 1, por lo que

$$u(x, y) = f(y + 2x) + g(y + 3x)$$

es una solución de la EDP conteniendo dos funciones arbitrarias f y g de sus respectivos argumentos. □

Acabamos esta sección clasificando las EDP de la forma (B.1). La clasificación de estas EDP surge por su analogía con la ecuación de las cónicas en el plano:

$$Ax^2 + Bxy + Cy^2 + Dx + Ey + F = 0.$$

Así, dependiendo de que la cantidad $B^2 - 4AC$ sea positiva, negativa o nula, hablaremos respectivamente de EDP hiperbólicas, elípticas o parabólicas.

Generalizando la definición anterior, decimos que la EDP (B.1) es de tipo

1. *hiperbólico* en todos los puntos en los que $B^2 - 4AC > 0$,
2. *elíptico* en todos los puntos en los que $B^2 - 4AC < 0$,
3. *parabólico* en todos los puntos en los que $B^2 - 4AC = 0$.

Esta clasificación puramente matemática se relaciona con una división global de los fenómenos físicos que se describen mediante tales ecuaciones, a saber: *procesos vibratorios* (ecuaciones hiperbólicas), *estacionarios* (ecuaciones elípticas), o de *difusión* (ecuaciones parabólicas). Es por ello que las soluciones de cada uno de los tipos de ecuaciones tienen particularidades que le son específicas (si bien pueden existir puntos de contacto). Por ejemplo, las ecuaciones hiperbólicas se caracterizan por poseer soluciones en forma de ondas que se desplazan con una velocidad finita. Las ecuaciones elípticas poseen soluciones suaves (infinitamente diferenciables), etc. Las ecuaciones parabólicas, en cierto sentido, tienen propiedades intermedias entre las hiperbólicas y las elípticas.

Ilustraremos esta clasificación con las tres EDP más famosas de la física matemática.

EJEMPLO. La ecuación $u_{xx} - u_{yy} = 0$ es hiperbólica puesto que $A = 1$, $B = 0$, $C = -1$ y $B^2 - 4AC = 4 > 0$. Es un caso especial de la denominada *ecuación unidimensional de ondas*, que es satisfecha por los pequeños desplazamientos transversales de los puntos de una cuerda vibrante. Esta EDP es lineal y homogénea con coeficientes constantes y posee la solución

$$u(x, y) = f(y + x) + g(y - x),$$

donde f y g son funciones arbitrarias de sus respectivos argumentos. □

Ejemplo. La ecuación $u_{xx} + u_{yy} = 0$ es elíptica puesto que $A = 1$, $B = 0$, $C = 1$ y $B^2 - 4AC = -4 < 0$. Esta EDP se denomina *ecuación bidimensional de Laplace* y es satisfecha por la temperatura de los puntos de una placa rectangular delgada en estado estacionario. Observemos que esta EDP es lineal homogénea con coeficientes constantes, por lo que su solución es

$$u(x, y) = f(y + ix) + g(y - ix),$$

donde f y g son funciones arbitrarias de sus respectivos argumentos. □

Ejemplo. La ecuación $u_{xx} - u_y = 0$ es parabólica puesto que $A = 1$, $B = C = 0$ y $B^2 - 4AC = 0$. Es un caso especial de la *ecuación unidimensional del calor* (o *ecuación de difusión*), que es satisfecha por la temperatura de un punto de una barra homogénea. Esta EDP no es homogénea. □

B.4. Forma canónica de las EDP lineales de segundo orden

La utilidad de la clasificación anterior se basa esencialmente en la posibilidad de reducir (B.1), en cada uno de los tres casos mencionados, a una *forma canónica* mediante un adecuado cambio de variables independientes.

Para determinar estas formas canónicas, empezamos considerando un cambio genérico de variables independientes:

$$s = s(x, y) \qquad \text{y} \qquad t = t(x, y),$$

donde supondremos que s y t son funciones de x e y dos veces derivables, de manera que el jacobiano de la transformación,

$$J = \begin{vmatrix} s_x & s_y \\ t_x & t_y \end{vmatrix},$$

es distinto de cero en la región en que estemos interesados. Entonces, suponiendo que x e y son a su vez funciones de s y t dos veces derivables, las derivadas que aparecen en (B.1) se transforman en

$$\begin{aligned}
u_x &= u_s s_x + u_t t_x \\
u_y &= u_s s_y + u_t t_y \\
u_{xx} &= u_{ss} s_x^2 + 2u_{st} s_x t_x + u_{tt} t_x^2 + u_s s_{xx} + u_t t_{xx} \\
u_{xy} &= u_{ss} s_x s_y + u_{st}(s_x t_y + s_y t_x) + u_{tt} t_x t_y + u_s s_{xy} + u_t t_{xy} \\
u_{yy} &= u_{ss} s_y^2 + 2u_{st} s_y t_y + u_{tt} t_y^2 + u_s s_{yy} + u_t t_{yy}
\end{aligned}$$

que al sustituir en (B.1) da

$$A_1 u_{ss} + B_1 u_{st} + C_1 u_{tt} + R(s, t, u, u_s, u_t) = 0, \tag{B.5}$$

donde R es una función lineal en u, u_t, u_s e independiente de las derivadas segundas. Los nuevos coeficientes son

$$\begin{aligned}
A_1 &= A s_x^2 + B s_x s_y + C s_y^2, \\
B_1 &= 2A s_x t_x + B(s_x t_y + s_y t_x) + 2C s_y t_y, \\
C_1 &= A t_x^2 + B t_x t_y + C t_y^2.
\end{aligned}$$

Obsérvese que la naturaleza de (B.1) permanece invariante ante la transformación efectuada, puesto que como puede comprobarse fácilmente

$$B_1^2 - 4A_1 C_1 = J^2(B^2 - 4AC),$$

y por tanto la ecuación, después del cambio, continúa perteneciendo a la misma clase, sin más que exigir que $J \neq 0$.

Ecuaciones hiperbólicas

En el caso concerniente a la forma canónica en el caso hiperbólico, si elegimos una transformación de coordenadas tal que $s(x,y)$ y $t(x,y)$ sean dos soluciones de

$$A\left(\frac{dy}{dx}\right)^2 - B\left(\frac{dy}{dx}\right) + C = 0;$$

es decir,

$$\frac{dy}{dx} = \frac{B+\sqrt{B^2-4AC}}{2A} \quad \text{y} \quad \frac{dy}{dx} = \frac{B-\sqrt{B^2-4AC}}{2A}, \tag{B.6}$$

entonces el cambio de variables $s = s(x,y)$ y $t = t(x,y)$ simplifica la EDP (B.5) como sigue

$$u_{st} = -\frac{R}{B_1},$$

que se denomina *primera forma canónica de las EDP hiperbólicas.* Las curvas descritas al hacer

$$s(x,y) = \text{constante} \quad \text{y} \quad t(x,y) = \text{constante}$$

se denominan **curvas características**. Un tipo de curva característica para una EDP dada es una curva sobre la cual la solución toma un valor constante.

Si introducimos nuevas variables w y z, mediante las fórmulas

$$w = \frac{s+t}{2} \quad \text{y} \quad z = \frac{s-t}{2},$$

obtenemos

$$u_{ww} - u_{zz} = -4\frac{R}{B_1},$$

que recibe el nombre de *segunda forma canónica de las EDP hiperbólicas.*

Ecuaciones elípticas

En este caso el procedimiento es el mismo que para las ecuaciones hiperbólicas, pero, como $B^2-4AC<0$, (B.6) tiene como soluciones dos funciones complejas conjugadas $s = \alpha + i\beta$ y $t = \alpha - i\beta = \bar{s}$. Siguiendo el mismo procedimiento que en el caso hiperbólico y, para no trabajar con cantidades complejas, considerando el cambio de variables

$$w = \frac{s+t}{2} = \alpha \quad \text{y} \quad z = \frac{s-t}{2i} = \beta,$$

llegamos a la *forma canónica de las EDP elípticas*

$$u_{ww} + u_{zz} = -\frac{4R}{B_1}.$$

Ecuaciones parabólicas

Como $B^2 - 4AC = 0$, de (B.6) vemos que se satisface la misma EDO

$$\frac{dy}{dx} = \frac{B}{2A}.$$

Las soluciones de esta EDO, $s(x,y) = \text{constante}$, son las características de la EDP (B.1). (Para las ecuaciones parabólicas solo existe una familia de curvas características.) Tomando $s = s(x,y)$, podemos tomar una función arbitraria $t = t(x,y)$ que complete el cambio de variables y tal que el jacobiano $J \neq 0$. Se suele tomar $t = y$. En resumen, la ecuación se reduce a la *forma canónica de las EDP parabólicas*

$$u_{tt} = -\frac{R}{C_1}.$$

Caso particular: A, B y C constantes

Si A, B y C son constantes, $B^2 - 4AC$ también es constante, y las EDO (B.6) se pueden escribir como

$$y'(x) = \frac{B - \sqrt{B^2 - 4AC}}{2A} = \lambda_1 \qquad \text{e} \qquad y'(x) = \frac{B + \sqrt{B^2 - 4AC}}{2A} = \lambda_2;$$

es decir, dos EDO de primer orden, cuyas soluciones se obtienen mediante integración directa

$$y(x) = \lambda_1 x + K_1 \qquad \text{e} \qquad y(x) = \lambda_2 x + K_2.$$

Como las dos curvas características dependen de una constante arbitraria, basta elegir s como una de ellas y t como la otra para obtener el cambio de variables

$$s = y - \lambda_1 x, \qquad t = y - \lambda_2 x.$$

Obsérvese que λ_1 y λ_2 son las raíces de la ecuación de segundo grado $A\lambda^2 - B\lambda + C = 0$.

1. *Caso hiperbólico:* hay dos familias de rectas características. El cambio de variabales es

$$s = y - \lambda_1 x \qquad \text{y} \qquad t = y - \lambda_2 x.$$

2. *Caso elíptico:* no hay características. El cambio de variabales es

$$s = \frac{2Ay - Bx}{\sqrt{4AC - B^2}}, \qquad t = x,$$

puesto que, como $s = y - \lambda_1 x$, se tiene que

$$s = \frac{2Ay - Bx}{2A} + i\,\frac{\sqrt{4AC - B^2}}{2A}\,x = \text{constante},$$

que es equivalente a

$$s = \frac{2Ay - Bx}{\sqrt{4AC - B^2}} + ix = \text{constante}.$$

3. *Caso parabólico:* hay una familia de rectas características. El cambio de variabales es

$$s = y - \frac{B}{2A}x, \qquad t = y.$$

En algunos casos será posible hallar elementalmente la solución general de la EDP (B.1) una vez escrita en su forma canónica, pero en la mayoría de los casos será imposible. Identifiquemos dos casos en los que si es posible:

i) Si solo aparecen derivadas respecto a una variable. Por ejemplo,

$$u_{tt} + E_1 u_t + F_1 u = G_1.$$

La EDP se integra considerando la otra variable como parámetro. La EDP es parabólica.

ii) Si solo aparecen u_{st} y una de las derivadas primeras. Por ejemplo,

$$u_{st} + D_1 u_s = G_1.$$

La EDP se resuelve haciendo el cambio $u_s = v$, ya que la ecuación resultante $v_t + D_1 v = G_1$ se puede integrar considerando s como parámetro. La EDP es hiperbólica.

Ejemplo. Sea la EDP

$$4u_{xx} + 5u_{xy} + u_{yy} + u_x + u_y = 2.$$

En primer lugar, vemos que $A = 4$, $B = 5$, $C = 1$ y $B^2 - 4AC = 9 > 0$, de modo que la ecuación es hiperbólica. Como A, B y C son constantes, consideramos la transformación $s = y - \lambda_1 x$ y $t = y - \lambda_2 x$,

donde λ_1 y λ_2 son las raíces de la ecuación de segundo grado $4\lambda^2 - 5\lambda + 1 = 0$. Luego, $\lambda_1 = \frac{1}{4}$ y $\lambda_2 = 1$, de manera que $s = y - \frac{x}{4}$ y $t = y - x$. Aplicando esta transformación a la EDP, obtenemos su forma canónica

$$u_{st} - \frac{u_s}{3} = -\frac{8}{9}.$$

Hacemos $u_s = v$ y la ecuación resultante, $v_t = \frac{v}{3} - \frac{8}{9}$, es una EDO separable de primer orden, manteniendo s constante. Por tanto, $v = \phi(s)\,\mathrm{e}^{\frac{t}{3}} + \frac{8}{3}$, donde ϕ es una función arbitraria de s. Integramos a continuación este resultado parcialmente respecto a s, manteniendo t constante, para obtener

$$u(s,t) = f(s)\,\mathrm{e}^{\frac{t}{3}} + \frac{8}{3}s + g(t),$$

donde f es una función arbitraria de s que está definida por $f(s) = \int \phi(s)\,ds$ y g es una función arbitraria de t. En consecuencia,

$$u(x,y) = f\left(y - \frac{x}{4}\right)\mathrm{e}^{\frac{1}{3}(y-x)} + \frac{8}{3}\left(y - \frac{x}{4}\right) + g(y-x),$$

donde f y g son dos funciones arbitrarias de sus respectivos argumentos. □

B.5. Separación de variables

Si una EDP con dos o más variables independientes puede reducirse a un conjunto de EDO, una para cada variable, la ecuación se dice **separable**. Las soluciones de la EDP son entonces los productos de las soluciones de las EDO.

En esta sección introducimos el *método de separación de variables*, que es un método fundamental y potente para obtener soluciones de ciertos problemas que implican EDP. Aunque la clase de problemas a los que se puede aplicar este método es relativamente limitada, incluye no obstante muchos casos de gran interés físico.

El desarrollo del método requiere conectar con dos clases importantes de problemas, que históricamente surgieron precisamente de él. Son los problemas de contorno con EDO y el problema de representación de una función en forma de serie trigonométrica. A dichas series trigonométricas se les denomina *series de Fourier*; su estudio se escapa de los objetivos de este texto, pero sus propiedades se pueden encontrar en cualquier texto sobre EDP.

El enunciado matemático de tales problemas contiene una EDP y ciertas condicones suplementarias (condiciones de contorno, condiciones iniciales o ambas), y la solución del problema es una función que satisface tanto la EDP como las condiciones suplementarias.

Si suponemos, por ejemplo, una EDP con una sola variable dependiente u que es una función de dos variables independientes x e y, la idea del método de separación de variables consiste en buscar una solución de la EDP lineal de orden dos en la forma $u(x,y) = X(x)Y(y)$. En el caso en que sea aplicable este método, su aplicación lleva tres pasos:

1. obtención de dos EDO,
2. obtención de las soluciones de las dos EDO que cumplan las condiciones de contorno,
3. formación de una combinación lineal infinita de las soluciones para satisfacer las condiciones iniciales del problema.

El paso 3 se consigue primeramente aplicando la siguiente **generalización del teorema fundamental** visto anteriormente:

> Sean $u_1, u_2, \ldots, u_n, \ldots$ una infinidad de soluciones de la EDP (B.2) en una región $\mathcal{R}$ del plano XY. Supongamos que la serie infinita $\sum_{n=1}^{\infty} u_n = u_1 + u_2 + \cdots + u_n + \cdots$ converge a u en $\mathcal{R}$ y que es derivable término a término en $\mathcal{R}$ para obtener las diversas derivadas (de u) que aparecen en la EDP (B.2). La función u (definida por $u = \sum_{n=1}^{\infty} u_n$) es también una solución de la EDP (B.2) en $\mathcal{R}$.

El *objetivo* es obtener una solución que sea una serie, aplicándose después las condiciones iniciales del problema.

Señalemos que el procedimiento así esquematizado es estrictamente *formal*. No vamos a hacer ningún intento para justificar dicho procedimiento. En un tratamiento riguroso se ha de demostrar que la «solución formal» obtenida satisface realmente tanto la EDP como las condiciones suplementarias y que la solución así justificada es la única solución del problema.

Ejemplo: el problema de difusión

Ilustraremos los principios esenciales del **método de separación de variables** considerando el mismo problema que consideró Fourier sobre la conducción de calor en una varilla unidimensional cuyos extremos se mantienen a la temperatura constante de 0°C y la distribución de temperatura inicial está dada por la función $f(x)$. El modelo matemático que lo rige es el siguiente problema de contorno con condición inicial:

$$u_t = \beta u_{xx}, \quad \text{para} \quad 0 < x < \ell \quad \text{y} \quad t > 0 \quad (\beta = \text{constante}),$$

bajo las siguientes condiciones suplementarias

$$u(0,t) = u(\ell,t) = 0, \quad t > 0 \quad \text{(condiciones de contorno)},$$

$$u(x,0) = f(x), \quad 0 < x < \ell \quad \text{(condición inicial)}.$$

La idea es buscar soluciones de la forma $u(x,t) = X(x)T(t)$, para lo que se ha de verificar la ecuación

$$X(x)T'(t) = \beta X''(x)T(t), \quad \text{o bien,} \quad \frac{X''(x)}{X(x)} = \frac{T'(t)}{\beta T(t)} = \lambda,$$

donde necesariamente λ ha de ser una constante, que se denomina *constante de separación*. Por tanto, fijando λ, se tienen las dos EDO

$$X''(x) = \lambda X(x) \qquad \text{y} \qquad T'(t) = \beta\lambda T(t).$$

Como $u(x,t) = X(x)T(t)$, las condiciones de contorno son

$$X(0)T(t) = 0 \qquad \text{y} \qquad X(\ell)T(t) = 0, \qquad t > 0,$$

de manera que $T(t) = 0$, para todo $t > 0$, lo que implica que $u(x,t) \equiv 0$, o bien

$$X(0) = X(\ell) = 0.$$

Ignorando la solución trivial, se combinan estas últimas condiciones de contorno con la EDO correspondiente a $X(x)$ para obtener el problema de contorno

$$X''(x) = \lambda X(x); \qquad X(0) = X(\ell) = 0, \tag{B.7}$$

donde λ puede ser cualquier constante.

Notemos que la función $X(x) \equiv 0$ es una solución para todo λ y, dependiendo de la elección de λ, ésta puede ser la única solución del problema de contorno (B.7). Así que si se busca una solución no trivial $u(x,t) = X(x)T(t)$ del problema original, primero se deben determinar aquellos valores de λ para los cuales el problema de contorno (B.7) tiene una solución no trivial. Dichos valores especiales de λ se llaman **valores propios**, y las correspondientes soluciones no triviales son las denominadas **funciones propias**.

Para resolver el problema de contorno (B.7), se empieza con la ecuación auxiliar $r^2 - \lambda = 0$ y se consideran tres casos:

Caso 1: $\lambda > 0$. Las raíces de la ecuación auxiliar son $r = \pm\sqrt{\lambda}$, de modo que la solución general de la EDO de (B.7) es

$$X(x) = C_1 \,\mathrm{e}^{\sqrt{\lambda}\,x} + C_2 \,\mathrm{e}^{-\sqrt{\lambda}\,x}.$$

Para determinar C_1 y C_2, recurrimos a las condiciones de contorno

$$X(0) = C_1 + C_2 = 0 \qquad \text{y} \qquad X(\ell) = C_1 \,\mathrm{e}^{\sqrt{\lambda}\,\ell} + C_2 \,\mathrm{e}^{-\sqrt{\lambda}\,\ell} = 0.$$

de forma que si $C_2 = -C_1$, tenemos que $C_1 \left(\mathrm{e}^{\sqrt{\lambda}\,\ell} - \mathrm{e}^{-\sqrt{\lambda}\,\ell}\right) = 0$ o $C_1 \left(\mathrm{e}^{2\sqrt{\lambda}\,\ell} - 1\right) = 0$. Como hemos supuesto que $\lambda > 0$, resulta que $\mathrm{e}^{2\sqrt{\lambda}\,\ell} - 1 > 0$. Por lo tanto, C_1 y, en consecuencia, C_2 son iguales a cero. Por consiguiente, *no* existe solución no trivial de (B.7) para $\lambda > 0$.

Caso 2: $\lambda = 0$. Aquí $r = 0$ es una raíz doble de la ecuación auxiliar y la solución general de la EDO es entonces

$$X(x) = C_1 + C_2 x.$$

Las condiciones de contorno dadas en (B.7) originan las ecuaciones $C_1 = 0$ y $C_1+C_2\ell = 0$, las cuales implican que $C_1 = C_2 = 0$. Consecuentemente, para $\lambda = 0$, *no* existe solución no trivial de (B.7).

Caso 3: $\lambda < 0$. Las raíces de la ecuación auxiliar son $r = \pm i\sqrt{-\lambda}$, de modo que la solución general de la EDO que aparece en (B.7) es

$$X(x) = C_1 \cos\sqrt{-\lambda}\, x + C_2 \operatorname{sen}\sqrt{-\lambda}\, x.$$

Ahora las condiciones de contorno de (B.7) dan lugar al sistema

$$C_1 = 0, \qquad C_1 \cos\sqrt{-\lambda}\, \ell + C_2 \operatorname{sen}\sqrt{-\lambda}\, \ell = 0.$$

Puesto que $C_1 = 0$, el sistema se reduce a resolver $C_2 \operatorname{sen}\sqrt{-\lambda}\,\ell = 0$. Por lo tanto, $\operatorname{sen}\sqrt{-\lambda}\,\ell = 0$ o $C_2 = 0$. Ahora bien

$$\operatorname{sen}\sqrt{-\lambda}\,\ell = 0 \qquad \Leftrightarrow \qquad \sqrt{-\lambda}\,\ell = n\pi, \text{ donde } n \text{ es un entero.}$$

Por consiguiente, (B.7) tiene una solución no trivial ($C_2 \neq 0$) cuando $\sqrt{-\lambda}\,\ell = n\pi$ o $\lambda = -\left(\frac{n\pi}{\ell}\right)^2$, $n \in \mathbb{N}$ Además, las soluciones no triviales (funciones propias) $X_n(x)$ correspondientes al valor propio $\lambda = -\left(\frac{n\pi}{\ell}\right)^2$ están dadas por

$$X_n(x) = a_n \operatorname{sen}\left(\frac{n\pi x}{\ell}\right), \quad \in \mathbb{N},$$

donde los valores a_n son constantes arbitrarias distintas de cero.

Una vez determinado $\lambda = -(n\pi/\ell)^2$, para algún entero positivo n, consideremos la segunda EDO $T'(t) = \beta\lambda T(t)$ con $\lambda = -\left(\frac{n\pi}{\ell}\right)^2$; es decir,

$$T'(t) + \beta\left(\frac{n\pi}{\ell}\right)^2 T(t) = 0,$$

cuya solución general, para cada $n \in \mathbb{N}$, es

$$T_n(t) = b_n\, \mathrm{e}^{-\beta\left(\frac{n\pi}{\ell}\right)^2 t}.$$

Combinando ahora esta solución con la anterior, para cada $n \in \mathbb{N}$, se obtiene la función

$$u_n(x,t) = X_n(x)T_n(t) = a_n \operatorname{sen}\left(\frac{n\pi x}{\ell}\right) b_n\, \mathrm{e}^{-\beta\left(\frac{n\pi}{\ell}\right)^2 t} = c_n\, \mathrm{e}^{-\beta\left(\frac{n\pi}{\ell}\right)^2 t} \operatorname{sen}\left(\frac{n\pi x}{\ell}\right),$$

donde c_n es una constante arbitraria.

Señalemos que cada una de estas funciones u_n satisface tanto la EDP como las dos condiciones de contorno para todos los valores de las constantes c_n (compruébese como ejercicio).

Hemos de intentar ahora satisfacer la condición inicial. En general, por sí sola, ninguna de las anteriores soluciones $u_n(x,t)$ satisfará la condición inicial. Por ejemplo, si aplicamos la condición inicial a una solución $u_n(x,t)$, hemos de tener

$$f(x) = u_n(x,0) = c_n \operatorname{sen}\left(\frac{n\pi x}{\ell}\right), \quad 0 < x < \ell,$$

donde n es un entero positivo, y esto es evidentemente imposible, a menos que f sea una función sinusoidal de la forma $K \operatorname{sen}\left(\frac{n\pi x}{\ell}\right)$, para algún entero positivo n.

¿Qué hemos de hacer entonces? Según la generalización del teorema fundamental visto anteriormente, suponiendo la convergencia apropiada, una serie infinita de soluciones de la EDP del problema original es también solución. Formamos entonces una serie infinita

$$\sum_{n=1}^{\infty} u_n(x,t) = \sum_{n=1}^{\infty} c_n\, \mathrm{e}^{-\beta\left(\frac{n\pi}{\ell}\right)^2 t} \operatorname{sen}\left(\frac{n\pi x}{\ell}\right)$$

de las soluciones $u_n(x,t)$, que, por la generalización del teorema fundamental y suponiendo la convergencia apropiada, nos asegura que la suma de esta serie es también una solución de la EDP. Simbolizando dicha suma por $u(x,t)$, escribimos

$$u(x,t) = \sum_{n=1}^{\infty} u_n(x,t) = \sum_{n=1}^{\infty} c_n\, \mathrm{e}^{-\beta\left(\frac{n\pi}{\ell}\right)^2 t} \operatorname{sen}\left(\frac{n\pi x}{\ell}\right). \tag{B.8}$$

Observamos que $u(0,t) = u(\ell,t) = 0$, $t > 0$. Suponiendo entonces la convergencia apropiada, la función $u(x,t)$, dada por (B.8), satisface la EDP y las condiciones de contorno.

Aplicamos ahora la condición inicial a la solución (B.8):

$$f(x) = u(x,0) = \sum_{n=1}^{\infty} c_n \operatorname{sen}\left(\frac{n\pi x}{\ell}\right), \quad 0 < x < \ell.$$

De esta manera, el problema original de conducción de calor en una varilla unidimensional se ha reducido al problema de determinar un desarrollo de $f(x)$ de la forma

$$f(x) = \sum_{n=1}^{\infty} c_n \operatorname{sen}\left(\frac{n\pi x}{\ell}\right). \tag{B.9}$$

Un desarrollo de este tipo se llama **serie senoidal de Fourier**. Si se eligen los c_n de manera que la igualdad anterior sea válida, entonces el desarrollo de $u(x,t)$, dado por (B.8), se denomina **solución formal** del problema de conducción de calor en una varilla unidimensional. La terоría de las series de Fourier prueba que tales c_n son de la forma:

$$c_n = \frac{2}{\ell}\int_0^{\ell} f(s) \operatorname{sen}\left(\frac{n\pi s}{\ell}\right) ds, \quad n = 1, 2, \ldots$$

Notemos que la solución (B.8) es formal, ya que en el proceso de su obtención hemos supuesto una convergencia que no hemos justificado. Si este desarrollo converge a una función con segundas derivadas parciales continuas, entonces la solución formal es una solución verdadera (genuina). Además, la solución es única.

Ejemplo. Encuéntrese la solución del siguiente problema de difusión

$$\begin{cases} u_t = 7u_{xx}, \quad 0 < x < \pi, \quad t > 0, \\ u(0,t) = u(\pi,t) = 0, \quad t > 0, \\ u(x,0) = 3\operatorname{sen} 2x - 6 \operatorname{sen} 5x, \quad 0 < x < \pi. \end{cases}$$

Observemos que $\beta = 7$ y $\ell = \pi$. Por tanto, se requiere solo determinar los valores de c_n incluidos en la fórmula (B.9). Esto es

$$u(x,0) = 3\operatorname{sen} 2x - 6\operatorname{sen} 5x = \sum_{n=1}^{\infty} c_n \operatorname{sen} nx.$$

Igualando los coeficientes de términos semejantes, se encuentra que

$$c_2 = 3, \qquad c_5 = -6 \qquad \text{y} \qquad c_n = 0, \text{ para todo } n \neq 2, 5.$$

Por consiguiente, de (B.8), se sigue que la solución del anterior problema de difusión es

$$\begin{aligned} u(x,t) &= c_2 \, \mathrm{e}^{-\beta\left(\frac{2\pi}{\ell}\right)^2 t} \operatorname{sen}\left(\frac{2\pi x}{\ell}\right) + c_5 \, \mathrm{e}^{-\beta\left(\frac{5\pi}{\ell}\right)^2 t} \operatorname{sen}\left(\frac{5\pi x}{\ell}\right) \\ &= 3\,\mathrm{e}^{-28t} \operatorname{sen} 2x - 6\,\mathrm{e}^{-175t} \operatorname{sen} 5x. \quad \square \end{aligned}$$

B.6. Ejercicios

1. Hállese una solución que contenga dos funciones arbitrarias para cada una de las siguientes EDP. Determínese, para cada una de ellas, si la EDP es hiperbólica, elíptica o parabólica.

   a) $u_{xx} - 4u_{xy} + 4u_{yy} = 0$, *c*) $u_{xx} + 2u_{xy} + 5u_{yy} = 0$, *e*) $u_{xx} + 6u_{xy} + 9u_{yy} = 0$,

   b) $2u_{xy} + 3u_{yy} = 0$, *d*) $u_{xx} + u_{xy} - 6u_{yy} = 0$, *f*) $2u_{xx} - 2u_{xy} + 5u_{yy} = 0$.

2. Determínense y dibújense las regiones del plano para las cuales las siguientes EDP son de tipo hiperbólico, elíptico o parabólico.

   a) $u_{xx} + xu_{xy} - x^2 y = 0$, *b*) $u_{xx} + xu_{xy} - yu_{yy} - xyu_x = 0$.

3. Transfórmense las siguientes EDP en su forma canónica y resuélvanse aquellas que sean posibles.

   a) $u_{xx} + 4u_{xy} - 5u_{yy} + 6u_x + 3u_y - 9u = 0,$

   b) $u_{xx} - 6u_{xy} + 9u_{yy} + 2u_x + 3u_y - u = 0,$

   c) $u_{xx} + 2u_{xy} + 5u_{yy} + u_x - 2u_y - 3u = 0,$

   d) $u_{xx} - 5u_{xy} + 6u_{yy} = 0,$

   e) $u_{xx} - 4u_{xy} + 4u_{yy} = 0,$

   f) $x^2u_{xx} + 2xyu_{xy} + y^2u_{yy} + xyu_x + y^2u_y = 0.$

4. Utilícese el método de separación de variables para resolver cada uno de los siguientes problemas

   a) $u_t = u_x; \quad u(x,0) = e^x + e^{-2x},$

   b) $u_t = u_x; \quad u(0,t) = e^{-3t} + e^{2t},$

   c) $u_t = u_x + u; \quad u(x,0) = 2e^{-x} - e^{2x},$

   d) $u_t = u_x - u; \quad u(0,t) = e^{-5t} + 2e^{-7t} - 14e^{13t}.$

5. Determínese si se puede utilizar el método de separación de variables para resolver las siguientes EDP.

   a) $tu_{tt} + u_x = 0,$ *b)* $tu_{xx} + xu_t = 0,$ *c)* $u_{xx} + (x-y)u_{yy} = 0,$ *d)* $u_{xx} + 2u_{xy} + u_y = 0.$

6. Resuélvanse los siguientes problemas

   a) $u_t = (1.71)\ u_{xx}; \quad \begin{cases} u(x,0) = \operatorname{sen}\dfrac{\pi x}{2} + 3\operatorname{sen}\dfrac{5\pi x}{2}, & 0 < x < 2, \\ u(0,t) = u(2,t) = 0, & t > 0. \end{cases}$

   b) $u_t = (1.14)\ u_{xx}; \quad \begin{cases} u(x,0) = \operatorname{sen}\dfrac{\pi x}{2} - 3\operatorname{sen} 2\pi x, & 0 < x < 2, \\ u(0,t) = u(2,t) = 0, & t > 0. \end{cases}$

7. *El problema de la cuerda vibrante* ([21]). Aplíquese el método de separación de variables para encontrar una solución formal del problema de la cuerda vibrante, que analiza las vibraciones transversales de una cuerda sujeta entre dos puntos, tal como una cuerda de guitarra o de piano. Consideremos una cuerda elástica tensa cuyos extremos están fijos al eje X en los puntos $x = 0$ y $x = \ell$. Supongamos que para cada x en el intervalo $0 < x < \ell$ la cuerda se desplaza en el plano XY, siendo conocido el desplazamiento a partir del eje X, dado por la función $f(x)$. Supongamos también que en $t = 0$ se abandona la cuerda, a partir de la posición inicial dada por $f(x)$, con una velocidad inicial en cada punto del intervalo $0 \le x \le \ell$ dada por $g(x)$. El movimiento de dicha cuerda se rige por el siguiente problema de contorno con condiciones iniciales:

$$\begin{cases} u_{tt} = \alpha^2\, u_{xx}, & 0 < x < \ell, \quad t > 0, \\ u(0,t) = u(\ell,t) = 0, & t \ge 0, \\ u(x,0) = f(x), & 0 \le x \le \ell, \\ u_t(x,0) = g(x), & 0 \le x \le \ell. \end{cases}$$

   Como caso particular del problema de la cuerda vibrante resuélvase el siguiente problema ([19]). Supongamos que la cuerda es tal que la constante $\alpha^2 = 4$ y que los extremos están fijos al eje X en $x = 0$ y $x = \pi$. La distancia al eje X en el intervalo $0 \le x \le \pi$ está dada por $f(x) = \operatorname{sen} 3x - 4\operatorname{sen} 10x$. La velocidad inicial en cada punto del intervalo $0 \le x \le \pi$ con que se abandona la cuerda a partir de la posición inicial es $g(x) = 2\operatorname{sen} 4x + \operatorname{sen} 6x$.

8. Resuélvanse los siguientes problemas

   a) $u_{tt} = 9\ u_{xx}; \quad \begin{cases} u(x,0) = 6\operatorname{sen} 2x + 2\operatorname{sen} 6x, & 0 \le x \le \pi, \\ u_t(x,0) = 11\operatorname{sen} 9x - 14\operatorname{sen} 15x, & 0 \le x \le \pi, \\ u(0,t) = u(\pi,t) = 0, & t > 0. \end{cases}$

$$b)\ u_{tt} = u_{xx};\quad \begin{cases} u(x,0) = \sqrt{1-\cos x}, \quad 0 \leq x \leq 2\pi, \\ u_t(x,0) = 0, \quad 0 \leq x \leq 2\pi, \\ u(0,t) = u(2\pi,t) = 0, \quad t > 0. \end{cases}$$

9. Aplíquese el método de separación de variables para resolver el problema

$$u_t = u_{xx} + u;\quad \begin{cases} u(x,0) = 3\,\mathrm{sen}(2\pi x) - 7\,\mathrm{sen}(4\pi x), \quad 0 < x < 10, \\ u(0,t) = u(10,t) = 0, \quad t > 0. \end{cases}$$

10. *La ecuación bidimensional de Laplace* ([26]). Aplíquese el método de separación de variables para encontrar una solución formal de la ecuación bidimensional de Laplace. En dos dimensiones, el potencial gravitatorio newtoniano y el potencial electrostático vienen descritos por la ecuación $u_{xx} + u_{yy} = 0$. La ecuación bidimensional del calor, $u_t = \beta(u_{xx} + u_{yy})$, también se reduce a ella para el caso estacionario, ya que $u_t = 0$. También es una ecuación que aparece en hidrodinámica y en elasticidad. En estas circunstancias, en vez de dar la distribución inicial de temperaturas $f(x)$ a lo largo de la barra, como se hizo en el problema de difusión, se da la distribución de temperaturas en la frontera de la región D donde se satisface la ecuación de Laplace. Encuéntrese la solución de la ecuación

$$u_{xx} + u_{yy} = 0$$

en la región $D = \{(x,y)/\ 0 \leq x \leq a,\ 0 \leq y \leq b\}$, con las cuatro condiciones de contorno

$$\begin{aligned} u(x,0) &= 0, \quad u(x,b) = f(x), \quad 0 \leq x \leq a, \\ u(0,y) &= 0, \quad u(a,y) = 0, \quad\quad 0 \leq y \leq b. \end{aligned}$$

Este problema se conoce como *problema de Dirichlet.*

11. Hállese el potencial en la placa metálica

$$D = \{(x,y)/\ 0 \leq x \leq a,\ 0 \leq y \leq b\}$$

con las condiciones de contorno

$$u(0,y) = 0, \quad u(a,y) = 0, \quad u(x,0) = 0, \quad u(x,b) = 2\,\mathrm{sen}\left(\frac{2\pi x}{a}\right).$$

12. La ecuación bidimensional del calor es de la forma

$$u_t = \beta(u_{xx} + u_{yy}).$$

Suponiendo que $u(x,y,t) = X(x)Y(y)T(t)$, obténganse EDO para X, Y y T. Hállense las soluciones que satisfagan las condiciones

$$u(0,y,t) = 0, \quad u(a,y,t) = 0, \quad u(x,0,t) = 0, \quad u(x,b,t) = 0.$$

13. Obténgase la ecuación bidimensional de Laplace en coordenadas polares.

Bibliografía básica

[1] A. Aubanell, A. Benseny y A. Delshams, Útiles básicos de cálculo numérico, Labor, Barcelona, 1993.

[2] R. B. Bhat y S. Chakraverty, Numerical analysis in engineering, Alpha Science, Oxford, 2007.

[3] R. L. Burden y J. D. Faires, Análisis numérico, 7ª edición, Thompson, Madrid, 2003.

[4] S. C. Chapra y R. P. Canale, Métodos numéricos para ingenieros, McGraw-Hill Interamericana, México, 2007.

[5] A. Cordero, J. L. Hueso, E. Martínez y J. R. Torregrosa, Problemas resueltos de métodos numéricos, Thomson, Madrid, 2006.

[6] V. Domínguez y M. L. Rapún, Matlab en cinco lecciones de numérico, Universidad Pública de Navarra, Pamplona, 2007.

[7] J. D. Faires y R. Burden, Métodos numéricos, 3ª edición, Thomson, Madrid, 2004.

[8] L. V. Fausett, Applied numerical analysis using Matlab, 2nd edition, Pearson Prentice Hall, New Jersey, 2008.

[9] I. A. García y S. Maza, Métodos numéricos, Edicions de la Universitat de Lleida, Lleida, 2009.

[10] F. García y A. Nevot, Métodos numéricos en forma de ejercicios resueltos, Universidad Pontificia Comillas de Madrid, Madrid, 1997.

[11] C. F. Gerald y P. O. Wheatley, Análisis numérico con aplicaciones, 6ª edición, Pearson Educación, México, 2000.

[12] A. Gilat y V. Subramaniam, Numerical methods for engineers and scientists: an introduction with applications using MatLab, John Wiley and Sons, New Jersey, 2008.

[13] P. Holoborodko, Applied mathematics and beyond: numerical methods. Disponible en `http://www.holoborodko.com/pavel/numerical-methods/numerical-%20integration/`

[14] J. A. Infante y J. M. Rey, Métodos numéricos, 3ª edición, Pirámide, Madrid, 2007.

[15] E. Isaacson y H. B. Keller, Analysis of numerical methods, 2nd edition, John Wiley & Sons, New York, 1990.

[16] A. Kharab y R. B. Guenther, An introduction to numerical methods: a Matlab approach, Chapman and Hall/CRC, Boca Ratón, 2006.

[17] D. Kincaid y W. Cheney, Análisis numérico: las mátemáticas del cálculo científico, Addison-Wesley Iberoamericana, Buenos Aires, 1994.

[18] J. H. Mathews y K. D. Fink, Métodos numéricos con MATLAB, 3ª edición, Prentice Hall, Madrid, 1999.

[19] R. K. Nagle y E. B. Saff, Fundamentos de ecuaciones diferenciales, Addison-Wesley Iberoamericana, Delaware, 1992.

[20] D. Prasad, An introduction to numerical analysis, Alpha Science, Oxford, 2006.

[21] S. L. Ross, Ecuaciones diferenciales, Reverté, Barcelona, 1979.

[22] A. Quarteroni, R. Sacco y F. Saleri, Numerical Mathematics, 2nd edition, Springer-Verlag, New York, 2007.

[23] A. Quarteroni y F. Saleri, Cálculo científico con MATLAB y Octave, Springer-Verlag Italia, Milano, 2006.

[24] J. M. Quesada, C. Sánchez, J. Jódar y J. Martínez, Análisis y métodos numéricos, Universidad de Jaén, Jaén, 2004.

[25] V. Ramírez, D. Barrera, M. Posadas y P. González, Cálculo numérico con Mathematica, Ariel, Barcelona, 2001.

[26] J. San Martín, V. Tomeo y I. Uña, Métodos matemáticos, Thomson, Madrid, 2005.

[27] J. M. Sanz-Serna, Diez lecciones de cálculo numérico, Universidad de Valladolid, Valladolid, 1998.

[28] E. Steiner, Matemáticas para las ciencias aplicadas, Reverté, Barcelona, 2005.

Bibliografía complementaria

- K. E. Atkinson, An introduction to numerical analysis, John Wiley & Sons, New York, 1989.
- J. C. Butcher, Numerical methods for ordinary differential equations, 2nd edition, John Wiley & Sons, Chichester, 2008.
- S. D. Conte y C. de Boor, Análisis numérico elemental: un enfoque algorítmico, McGraw-Hill, New York, 1974.
- G. Dahlquist y Å. Björck, Numerical methods, Prentice-Hall, New Jersey, 1974.
- P. J. Davis, Interpolation and approximation, Dover, New York, 1975.
- P. J. Davis y P. Rabinowitz, Methods of numerical integration, 2nd edition, Dover, New York, 2007.
- A. J. Davis, The finite element method: an introduction with partial differential equations, Oxford University Press, Oxford, 2011.
- C. De Boer, A practical guide to splines, revised edition, Springer-Verlag, New York, 2001.
- P. Deuflhard y A. Hohmann, Numerical analysis in modern scientific computation: an introduction, 2nd edition, Springer-Verlag, New York, 2003.
- J. F. Epperson, An introduction to numerical methods and analysis, John Wiley & Sons, New Jersey, 2007.
- L. Fox, Numerical solution of two-point boundary value problem in ordinary differential equations, Dover, New York, 1990.
- M. Gasca, Cálculo numérico I, Publicaciones de la UNED, Madrid, 1991.
- W. Gautschi, Numerical analysis: an introduction, Birkhäuser, Boston, 1997.
- C. W. Gear, Numerical initial value problems in ordinary differential equations, Prentice Hall, Englewood Cliffs, NJ, 1971.
- C. F. Gerald y P. O. Wheatley, Applied numerical analysis, Addison-Wesley, New York, 1994.
- M. S. Gockenbach, Partial differential equations: analytical and numerical methods, SIAM, Philadelphia, 2011.
- G. H. Golub y J. M. Ortega, Scientific computing and differential equations: an introduction to numerical methods, Academic Press, London, 1992.
- G. H. Golub y C. F. Van Loan, Matrix computations, 3rd edition, Johns Hopkins University Press, Baltimore, 1996.
- C. Grossmann, H.-G. Roos y M. Stynes, Numerical treatement of partial differential equations, Springer-Verlag, Berlin Heidelberg, 2007.
- W. W. Hager, Applied numerical linear algebra, Prentice Hall, Englewood Cliffs, NJ, 1988.
- G. Hämmerling y K.-H. Hoffmann, Numerical mathematics, Springer-Verlag, New York, 1991.

- N. J. Higham, Accuracy and stability of numerical methods, SIAM, Philadelphia, 2002.
- H. Keller, Numerical methods for two-point boundary value problems, Dover, New York, 1992.
- J. D. Lambert, Numerical methods for ordinary differential systems: the initial value problem, John Wiley & Sons, Chichester, 1991.
- P. Lancaster y K. Salkaukas, Curve and surface fitting: an introduction, Academic Press, Boston, 1986.
- S. Larsson y V. Thomée, Partial differential equations with numerical methods, Springer-Verlag, Berlin Heidelberg, 2003.
- G. G. Lorentz, Approximation of functions, American Mathematical Society, Providence, 2005.
- J. M. Ortega y W. C. Reinboldt, Iterative solution of nonlinear equations in several variables, SIAM, Philadelphia, 2000.
- A. Ralston y P. Rabinowitz, A first course in numerical analysis, 2nd edition, Dover, New York, 2001.
- J. Stoer y R. Bulirsch, Introduction to numerical analysis, 3rd edition, Springer-Verlag, New York, 2002.
- C. W. Ueberhuber, Numerical computation: methods, software, and analysis, Springer-Verlag, Berlin Heidelberg, 1997.
- R. S. Varga, Matrix iterative analysis, 2nd edition, Springer-Verlag, Berlin Heidelberg, 2000.
- J. H. Wilkinson, Rounding errors in algebraic process, Dover, New York, 1994.
- Mathematica documentation center. Disponible en `http://reference.wolfram.com`.
- Mathworks documentation center. Disponible en `http://www.mathworks.es/help/documentation-center.html`.



Servicio de Publicaciones
Biblioteca Universitaria
C/ Piscinas, s/n
26006 Logroño (La Rioja)
Teléfono: 941 299 187

http://publicaciones.unirioja.es
www.unirioja.es